REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY *(Leave blank)*

2. REPORT DATE
14.Jan.03

3. REPORT TYPE AND DATES COVERED
DISSERTATION

4. TITLE AND SUBTITLE
A UNIFIED APPROACH TO STATISTICAL ASSESSMENT OF HEURISTIC QUALITY IN COMBINATORIAL OPTIMIZATION

5. FUNDING NUMBERS

6. AUTHOR(S)
CAPT GIDDINGS ANGELA P

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
PURDUE UNIVERSITY

8. PERFORMING ORGANIZATION REPORT NUMBER
CI02-805

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
THE DEPARTMENT OF THE AIR FORCE
AFIT/CIA, BLDG 125
2950 P STREET
WPAFB OH 45433

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION AVAILABILITY STATEMENT
Unlimited distribution
In Accordance With AFI 35-205/AFIT Sup 1

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

DISTRIBUTION STATEMENT A
Approved for Public Release
Distribution Unlimited

20030225 102

14. SUBJECT TERMS

15. NUMBER OF PAGES
146

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT

18. SECURITY CLASSIFICATION OF THIS PAGE

19. SECURITY CLASSIFICATION OF ABSTRACT

20. LIMITATION OF ABSTRACT

I dedicate this thesis to Jesus Christ, my Lord and Savior, and to Scott, Aaron, and AJ. Without the loving support of all of them, none of this would have been possible or worthwhile.

"I can do all things through Jesus Christ who strengthens me." Philippians 4:13
"[I]n all your ways acknowledge him, and he will make your paths straight." Proverbs 3:6

Graduate School Form 9
(Revised 7/94)

PURDUE UNIVERSITY
GRADUATE SCHOOL
Thesis Acceptance

This is to certify that the thesis prepared

By Angela P. Giddings

Entitled A Unified Approach to Statistical Quality Assessment in Heuristic Combinatorial Optimization

Complies with University regulations and meets the standards of the Graduate School for originality and quality

For the degree of Doctor of Philosophy

Signed by the final examining committee:

_______________, chair

Bruce W. Schmeiser co-chair

Julie Ann Stuart

John Deely

Approved by: Dennis Engi 8/22/2002
Department Head Date

This thesis ☐ is ☒ is not to be regarded as confidential.

Major Professor

Format Approved by:

Chair, Final Examining Committee or Thesis Format Adviser

ACKNOWLEDGMENTS

I would like to express my sincere appreciation to all the members of my research committee, Bruce Schmeiser, Reha Uzsoy, Ron Rardin, John Deely, and Julie Ann Stuart. Thank you all for advising me, educating me, and mentoring me with unfaltering patience and enthusiasm. I am exceedingly proud to say that I collaborated with such an amazing group.

I would also like to thank the other folks who encouraged me to take on this incredible academic journey toward a PhD. This list includes but is not limited to my parents, Scott, Col Mike Schiefer, Glenn Bailey, Bruce Merrill, and Col Jerry Diaz.

TABLE OF CONTENTS

Page

Page

Page

LIST OF FIGURES

Figure Page

ABSTRACT

Giddings, Angela P. Ph.D., Purdue University, December 2002. A Unified Approach to Statistical Assessment of Heuristic Quality in Combinatorial Optimization. Major Professors: Reha Uzsoy and Bruce Schmeiser.

Since the introduction of mathematical programming it has been all too easy to identify real-world problems that could be formulated as math programs but could not be solved to a provable optimum within a reasonable amount of time. As computing power continues to increase, so too does the size of the mathematical programs to be solved. This situation has given rise to a multitude of heuristic solution techniques that seek to provide good approximate solutions within a reasonable amount of time.

Designers and users of heuristic solution techniques would like to assess the quality of their heuristics, where heuristic quality is defined in terms of the characteristics of the solutions returned by the heuristic, often emphasizing the objective function values. Fixed bounds on worst case performance are available for some heuristics, but in many cases heuristic-quality assessment approaches must take a sampling perspective and apply statistical tools to derive their assessment. Although many authors have proposed statistical methods for assessing heuristic quality, there has not been a foundation for a single unified approach or a framework for comparison of the distinct approaches to heuristic-quality assessment.

The primary contribution of this research is that it presents a unifying probability modeling framework that applies whenever randomized heuristic solution techniques are applied to instances of combinatorial optimization problems. With this probability model in hand, we can better understand the relative strengths and weaknesses of the existing statistical approaches to assessing heuristic quality in combinatorial optimization. Moreover, the probability model suggests new avenues for the development of heuristic-quality assessment approaches, and we present empirical results from initial applications.

1 INTRODUCTION

1.1 Overview

As heuristic techniques are applied to combinatorial optimization problems with increasing frequency, users and developers of these techniques continue to search for an appropriate means of assessing the quality of the heuristics and the solutions they provide. Although deterministic lower bounds are available for some problem instances and worst-case performance limits are available for some heuristics, these assessments may be unsatisfactory as a reflection of expected performance of good heuristics on average problem instances. As a result, statistical methods have been proposed for assessing heuristic quality. This paper will examine previous statistical approaches to assessing heuristic quality and propose a new approach based upon a unifying probability modeling framework.

We begin our discussion by laying out the context of the heuristic-quality assessment in the remainder of Chapter 1. We explain the basic experiment that occurs when a randomized heuristic is used to solve combinatorial optimization problems. In this experiment, the heuristic is applied to a particular problem instance so that each application produces a single feasible solution and its associated objective function value, or *solution value*. In Chapter 2, we specify a new probability modeling framework for the solution value. We present a taxonomy of the different objectives and statistical paradigms in the literature. In Chapter 3, we discuss the heuristic-quality assessment literature in light of our new modeling framework and the associated taxonomy. A new heuristic-quality assessment approach is presented in Chapter 4, with an empirical investigation of this approach in Chapter 5. In Chapter 6 we summarize the strengths and weaknesses of the new approach and suggest areas for future research.

1.2 Math-Programming Formulation

Since statistical heuristic-quality assessment lies at the juncture of two very different operations research sub-disciplines—mathematical programming and statistical methods—we begin by defining terminology and notation.

Although many of the concepts we present apply equally well to nonlinear optimization problems and problems with continuous unbounded decision variables, we focus our discussion on (mixed) integer linear programs with bounded variables and minimization objective. A feasible solution to the optimization problem is represented by the vector $\mathbf{x}$ whose components contain fixed values for each of the decision variables in the problem. The solution value associated with $\mathbf{x}$ is $z(\mathbf{x})$, or simply z when only solution values are discussed. We assume that there are only $\psi < \infty$ feasible solution vectors to consider. This may require the addition of upper bound constraints on some of the variables along with a focus on extreme points with respect to the continuous-valued variables.

1.3 Terminology

In this section we discuss the two structures that come into play when heuristic solution techniques are applied to optimization problems—the problem class structure and the heuristic class structure. Although it has not been clearly stated in previous literature, heuristic-quality assessment is very much an attempt to understand and isolate the effects of these two classes on our observed solutions and solution values.

1.3.1 Problem Class

The term *problem instance* may be familiar to readers from the math-programming community, but we use this term and the term *problem class* in a fashion that differs in subtle but important ways from what most readers are used to. A *problem instance*, $\mathbb{I}$, is a particular math-programming problem with fixed values specified for its data (i.e. the constraint matrix and objective function coefficients). A *problem class*, $\mathcal{I}$, is a group of related *problem instances*, so $\mathbb{I} \in \mathcal{I}$. In the mathematical-programming literature, we often

think of a problem class as a very broad category of optimization problems such as traveling salesman problems, quadratic assignment problems, knapsack problems, or single-machine scheduling problems. Although our definition covers that type of problem class, it also covers much smaller collections such as a group of problems that are identical in all except the value of one element of problem data or even a problem class that contains precisely one problem instance, $\mathcal{I} = \{\mathbb{I}\}$.

The problem instance may be viewed as a single member of some problem class, selected for more detailed investigation, perhaps through some means of random sampling. If the problem class is in fact defined by computer code for problem generation, then the problem instance is precisely the combination of the problem class computer code with a random number seed specifying a sequence of random numbers to be used by the problem generation code.

1.3.2 Heuristic Class

Although the idea of a *heuristic class* is probably new to all of our readers, the concept is structurally similar to the definition of problem class given above. A *heuristic realization*, $\mathbb{H}$, is the heuristic procedure (e.g. general computer instructions for the heuristic) plus a specified random number seed. This heuristic realization is contained in a larger group which we call the *heuristic class*, $\mathcal{H}$, so $\mathbb{H} \in \mathcal{H}$. The *heuristic class* is the collection of all possible heuristic realizations for this procedure (i.e. the procedure plus the all possible random number seeds). This breakdown is most relevant to randomized heuristics but it also applies in a trivial fashion to deterministic heuristics, which are simply heuristic classes consisting of a single heuristic realization.

Technically, no heuristic procedure exists independent of some general problem class, so we could bring in a subscript on $\mathcal{H}$ reflecting the most general problem class to which it may apply. For the sake of notational simplicity we omit this subscript with the understanding that we only discuss heuristic procedures relative to problem classes when they may be applied to all instances from that class.

1.4 Notation

We use the convention of representing random variables in upper case while lower case is used for fixed values such as sample observations. Bold type is used when naming a vector or matrix. When sets of items are discussed, parentheses index items in observational order with brackets used when items are ordered by magnitude. For example, $\{z[1], z[2], \ldots z[k]\}$ are fixed values listed so that $z[1] \le z[2] \le \cdots \le z[k]$ while $\{w(1), w(2), \ldots, w(m)\}$ are fixed values listed in order of observation. For clarity, the ith order statistic of a sample of size n is designated with $[i, n]$ as in $y[i, n]$.

Parameters of probability distributions are generally named in Greek letters with their point estimates set off using a hat. For example, θ is our symbol for any distributional parameters that correspond to the objective function value at the optimal solution, or *optimal value*. A point estimate of θ is designated $\hat{\theta}$.

1.5 The Experiment behind Heuristic Solution of Combinatorial Optimization

When a practitioner applies a heuristic to solution of combinatorial optimization problems he is interested in a problem class, $\mathcal{I}$, and heuristic class, $\mathcal{H}$. In fact, the simple expression $\mathcal{I}\times\mathcal{H}$ captures how the heuristic practitioner begins by selecting $\mathbb{I}$ from $\mathcal{I}$ then selecting $\mathbb{H}$ from $\mathcal{H}$ for a single application to $\mathbb{I}$. Often the same heuristic class is applied n independent times to a single problem instance $\mathbb{I}$. Using distinct random number seeds for these n replications results in n different heuristic realizations. In this case the full experiment is $\mathcal{I}\times\mathcal{H}^n$.

What are the random variables associated with this experiment? Although the solution values, $\{z[1], z[2], \ldots, z[\psi]\}$, and their associated solution vectors, $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_\psi\}$, are fixed for a particular problem instance $\mathbb{I}$, they may be considered random variables with respect to the problem class $\mathcal{I}$, particularly if the full experiment involves multiple problem

instances as in $(\mathcal{I}\times\mathcal{H}^n)^m$. Even when the problem instance is fixed, the solution values and solution vectors *returned by the heuristic* are random variables. We denote these by W and $\mathbf{Y}$, respectively.

1.6 Coin-Flip Analogy

At the heart of any heuristic-quality assessment approach is the interaction of $\mathbb{I}$ and $\mathbb{H}$ to produce $(\{z[1], z[2], \ldots, z[\psi]\},\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_\psi\}; w, \mathbf{y})$, where $(\{z[1], z[2], \ldots, z[\psi]\},\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_\psi\})$ reflects only the problem instance while $(w, \mathbf{y})$ is a single observation resulting when the heuristic is applied to the problem instance. This is the simple component experiment in heuristic-quality assessment in much the same way that a single coin flip is the component Bernoulli experiment of the prototypical coin flipping binomial experiment. Since it will provide a concrete mental image for our discussion of heuristic-quality assessment, let's take the time to expand on the coin-flip analogy.

The problem dimension in heuristic-quality assessment corresponds to the coin dimension. Selecting a particular problem instance is like selecting a particular coin to flip from some container of coins, which corresponds to the problem class. As the definition of a problem class may vary greatly, we may likewise select our coin from a container that contains a small group of virtually identical coins or from a huge collection of coins of all conceivable denominations.

The heuristic dimension in heuristic-quality assessment corresponds to the flip dimension. Just as there are a wide variety of methods for flipping a coin, there are a wide variety of heuristic procedures, or $\mathcal{H}$'s. As we limit ourselves to consideration of only heuristic procedures that are applicable to every member of the considered problem class, we could only consider coin flipping procedures if they were practical for every coin in the collection from which we will select our coin. Even upon specification of a particular flipping technique, each flip applying that technique may have a random component that produces widely varying results, or a highly precise procedure might be defined to have the same result or nearly the same result every time. Similarly, randomized heuristic

procedures might correspond to a wide variety of distinct $\mathbb{H}$'s or $\mathcal{H}$ might be defined as precisely the singleton $\{\mathbb{H}\}$, as is the case for non-randomized constructive heuristics (deterministic heuristics).

There are analogies with respect to the type of probability statements that might be made as well. In the coin-flip example, we might be interested in comparing the performance of particular flipping techniques over a collection of coins. This case corresponds to a focus on the performance of a heuristic or heuristics over a problem class. Here we seek probability statements involving W, perhaps conditioned on $\mathbb{I}$ or on some subset of $\mathcal{I}$. On the other hand, we might be more interested in the characteristics of coins in a given collection. Although we might use flipping techniques to gather information about the coins, we seek information about the coins themselves rather than about particular flipping approaches. This case corresponds to a focus on a problem class rather than on a heuristic procedure. Here the goal is some type of probability statement on Z, averaged over $\mathcal{H}$ where this heuristic procedure is the union of all applicable heuristic procedures for $\mathcal{I}$.

1.7 Different Perspectives in Heuristic-Quality Assessment

Many approaches have been proposed for statistical assessment of heuristic quality in combinatorial optimization. These approaches reflect differing perspectives in terms of the type of practitioner who is expected to use the quality assessment, the reason for the quality assessment, and the property of $\mathcal{I}\times\mathcal{H}$ that is the focus of the quality assessment. Sections 1.7.1 through 1.7.3 delineate these differences in perspective.

1.7.1 Type of Practitioner: Heuristic User or Heuristic Researcher

The nature of the desired probability statement is heavily influenced by the perspective of the person performing the heuristic-quality assessment, who we generically

call the *practitioner*. A *heuristic user* seeks to measure the quality of a heuristic solution to a particular optimization problem instance, $\mathbb{I}$. In our coin-flip analogy, this corresponds to an interest in a particular coin, with coin flips used merely as a means of gathering data about the coin. In contrast, a *heuristic researcher* wants to make quality statements for a proposed heuristic class, $\mathcal{H}$, over a fairly large problem class, $\mathcal{I}$. In the coin-flip analogy, this corresponds to investigation of a particular flipping technique across some group of coins.

Unfortunately, information is generally not available at the desired level. For instance, a heuristic user may have prior knowledge of how a lower bound procedure and sampling heuristic perform on similar problem instances but not for the particular instance at hand. In this case he may use deductive reasoning to suppose that his prior information would be relevant to this particular instance. On the other hand, a heuristic researcher will investigate the behavior of his heuristic over some set of problem instances and try to use inductive reasoning to make conclusions about an entire problem class. This is the approach followed by much of existing literature that attempts to evaluate the performance of heuristics through computational experiments, e.g. Rardin and Uzsoy (2001). The situation is illustrated in Figure 1.

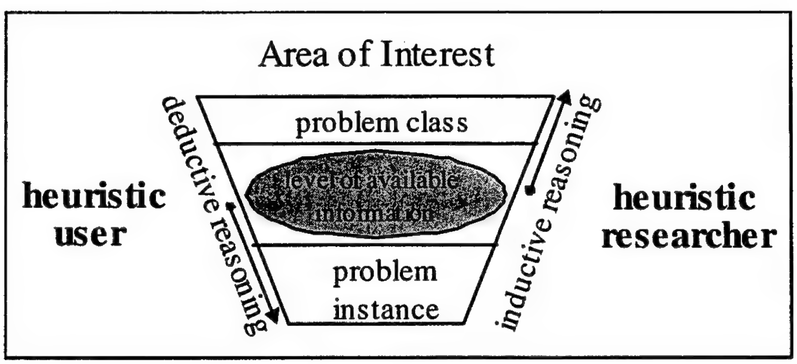

Figure 1. Comparing Heuristic User and Heuristic Researcher Perspectives

1.7.2 Objective: Estimation or Stopping-Rule Development

Clearly, the heuristic-quality assessment pursued by a heuristic user is different from that sought by a heuristic researcher. Moreover, any practitioner may have one or

both of the following goals in assessing heuristic quality. The first goal is estimating some property of the problem instance and/or heuristic class that provides a reflection of solution quality. The second goal is determining when enough heuristic realizations have been collected, and he should stop sampling. A stopping rule often incorporates an estimate of some property of the problem instance or heuristic. Similarly, an estimation approach can be used along with a cost or utility structure to design a stopping rule.

Both the estimation problem and the stopping-rule problem can occur at either the problem-instance or the problem-class level. For example, a heuristic user needs a stopping rule to know when to stop applying his randomized heuristic to a particular problem instance. A heuristic researcher might also need a stopping rule for application of a randomized heuristic to an individual problem instance. Moreover, a heuristic researcher could need a stopping rule to help him determine when enough problem instances have been explored through the application of randomized or deterministic heuristics. Similarly, statistical estimation can be used for inference at the problem instance level by the heuristic user and on the problem class/heuristic class level by the heuristic researcher.

1.7.3 Focus: Number of Local Optima or Optimal Value

There is yet another difference of perspective among those interested in assessing heuristic quality. Some authors focus on estimating the number of locally optimal solutions to the particular problem instance. Others focus on the optimal value θ, which is the objective function value corresponding to the optimal solution of a particular problem instance.

As Section 1.7.1 through Section 1.7.3 illustrate, many distinct perspectives exist in heuristic-quality assessment. However, the basic probability structure underlying heuristic solution of optimization can unify all these approaches. All existing literature on statistical heuristic-quality assessment for combinatorial optimization can be viewed as attempts to estimate or develop a stopping rule based upon some property of the probability model presented in the next section or a simplified version of this model.

2 A UNIFYING FRAMEWORK FOR STATISTICAL HEURISTIC-QUALITY ASSESSMENT

2.1 The P-H Probability Model

We focus our development on a model that is most directly relevant to estimating the number of local optima or the optimal value from a heuristic user's perspective (i.e. for a single problem instance). This perspective is the simplest combination of those listed in Section 1.7, yet a full development from this perspective has clear implications for heuristic researchers, inference on the number of local optima, and stopping rule development.

Our context here is the $\mathcal{I}\times\mathcal{H}^n$ experiment. When the focus is on estimating the number of local optima, we need a probability model for the observation (w, $\mathbf{y}$) that has ψ as a parameter. In optimal-value estimation, we need a probability model for the observation (w, $\mathbf{y}$) with the optimal value θ as a parameter. Since $\theta = z[1]$ is itself a solution value, it makes sense to focus on the solution value random variable, W, for optimal-value estimation, only using the observed solution vectors $\mathbf{y}(i)$ to indicate whether or not two observations with equivalent solution values correspond to the same solution, i.e. the same index on $\{1, 2, \ldots, \psi\}$. This focus on W is also appropriate for estimation of the number of optima, although a focus on $\mathbf{Y}$ could work as well.

We call our probability model the P-H Probability Model, since it is the first probability model proposed for heuristic-quality assessment that distinctly incorporates the structure of both the problem and heuristic classes.

> The *P-H Probability Model* is given by:
>
> $P(W = w) \equiv f_W(w) = \sum_{i \ni w = z[i]} p_i = \sum_{i=1}^{\psi} p_i \cdot \Delta(z[i] = w)$, where Δ is an indicator function.

This probability model is fairly intuitive. The probability of our heuristic returning a solution value of w is a sum of the probability that our heuristic returns any of the particular solution vectors that have solution value equal to w. Proper use of the P-H

Probability Model requires enough prior information about the heuristic to specify an appropriate structure for p_i, the probability that our heuristic returns the solution with the *i*th smallest solution value. Moreover, having an appropriate problem class model for the location of each $z[i]$ can further improve estimation. In some cases the practitioner may have little external information about the problem instance, problem class, or heuristic. When this is true, the details of the model can be difficult to specify.

2.2 Simplifications

The P-H Probability Model provides a unique lens for viewing the efforts of previous authors in heuristic-quality assessment. This is because all previous attempts to statistically assess heuristic quality can be viewed as various simplifications or special cases of the P-H modeling framework. Two particular simplifications are prevalent: ignoring the magnitude of solution values and using a continuous probability model for W. Sections 2.2.1 and 2.2.2 discuss these simplifications in more detail.

2.2.1 Ignoring the Magnitude of Solution Values

This simplification is at the heart of approaches that estimate the number of local optima rather than the optimal value. Ignoring the order in solution values simplifies the estimation problem in that when $(w(j), \mathbf{y}(j))$ is observed it is not necessary to determine its position in the ordering of previous observations or relative to the underlying solution space $\{(z[i], \mathbf{x}[i]), i=1,\ldots, \psi\}$. Instead, the practitioner simply determines whether $(w(j), \mathbf{y}(j))$ corresponds to a previously observed solution and increments the relevant counters. The drawback to this simplification is that it generally does not reflect the true priorities in heuristic solution of optimization problems. Knowing that he had observed nearly all of the existing local optima (i.e. nearly all i such that $p_i>0$) might convince the practitioner that he is unlikely to find a better optimum through more heuristic runs, but he might still be worried about how much better any remaining optima could be. Moreover, the practitioner would often prefer to avoid a large number of local optima whose solution values are dominated by other local optima rather than continuing to make heuristic runs until he estimates that there is a high probability that he has seen all of them.

2.2.2 Using a Continuous Probability Model for W

This simplification is common in optimal-value estimation approaches. Since ψ is often very large, it is tempting to argue that using a continuous probability model is acceptable due to the large number of solutions and the distinct solution values that presumably accompany these solutions. Clearly, the estimation problem is simpler if θ is the lower bound parameter of a continuous W probability distribution. However, this simplification also has its drawbacks.

First, the assumption of a continuous probability model on W with a lower bound at θ results in a model with zero probability of observing duplicate values and, in fact, zero probability of ever actually observing θ, since any continuous probability model has zero probability mass at any individual value. In essence, this simplification relies on the heuristic being random enough and the problem instance containing enough distinct solution values that the practitioner never detects the discreteness that lies beneath the simplifying assumption. If duplicate solutions are observed, he is forced to omit them from consideration in order to maintain the assumed "pseudo-distribution" model (Los and Lardinois 1982).

Another drawback to this approach is that assuming a single-stage continuous probability distribution for W confounds the effects of heuristic and problem on W, making it impossible to isolate which aspects of the W distribution reflect properties of the problem class and which reflect heuristic performance. This confounding of heuristic and problem effects calls into question attempts to assess heuristic performance across a problem class through repeated solution of the heuristic user's problem across multiple instances. Similarly, using a single-stage continuous probability model for W prevents the practitioner from effectively applying one heuristic class then transferring what is revealed about the problem instance when he applies a second heuristic class.

When considered together, these drawbacks indicate that using a continuous probability model for W makes a practitioner particularly susceptible to modeling error. Although he may invoke statistical goodness-of-fit tests to allay his fears, failure to reject the continuous probability model merely reflects a relative lack of information in the data, since the continuous model is fundamentally incorrect.

In some cases a continuous approximation may be required out of computational necessity. In these cases, the heuristic practitioner must be particularly cautious and treat his statistical results as best-case bounds due to the potential for high levels of undetectable modeling error.

Now that we have introduced the relevant terminology, the P-H Probability Model, and prevalent simplifications of it; we will discuss the existing literature on assessing heuristic quality in Chapter 3. Using the P-H Probability Model as our lens allows us to recognize the relationship among distinct threads in the previous literature and to assess the relative strengths and weaknesses of previous approaches in a concrete theoretical fashion rather than through computational experiments. Analysis of previous heuristic-quality assessment approaches like that in Chapter 3 was not possible before introduction of the P-H Probability Model.

3 PREVIOUS LITERATURE ON STATISTICAL HEURISTIC-QUALITY ASSESSMENT

In this chapter we review the existing literature on statistical assessment of heuristic quality. As suggested by the discussion in Sections 1.7 and 2.2, quite a variety of statistical approaches to assessing heuristic quality have appeared in a wide array of journals. Figure 2 shows the previous literature on heuristic-quality assessment, categorized according to the discussion of different perspectives and simplifications in Sections 1.7 and 2.2.

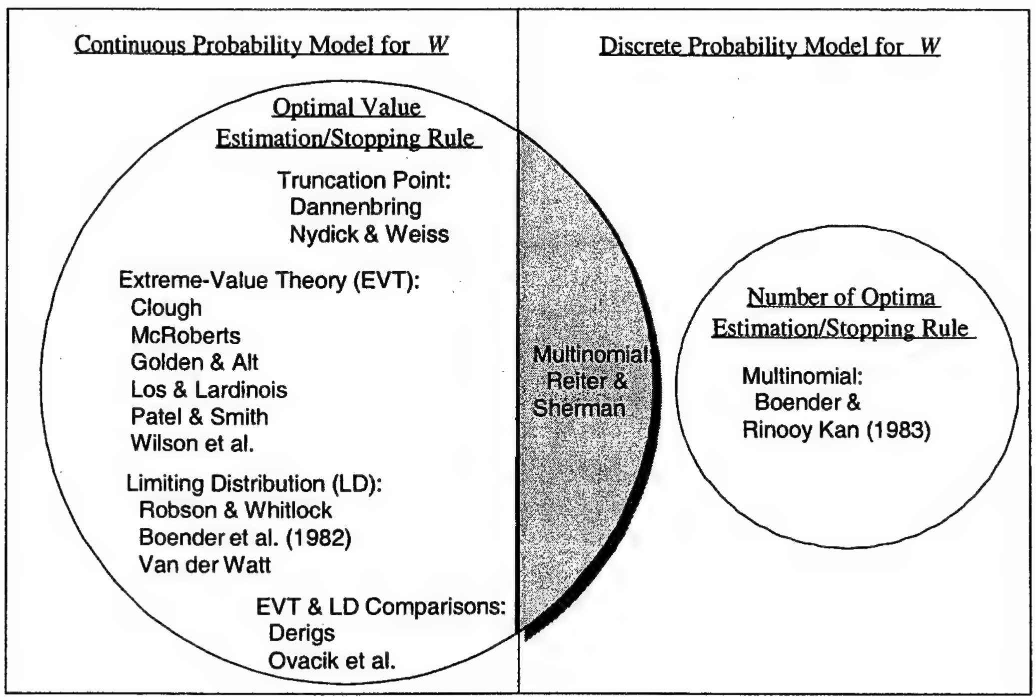

Figure 2. Previous Literature on Heuristic-Quality Assessment.

Our presentation of previous approaches proceeds from left to right across Figure 2. We begin in Section 3.1 with continuous optimal-value estimation approaches, which are most common in literature. In Section 3.1 we also discuss the single discrete approach to optimal-value estimation. We finish in Section 3.2 by covering the discrete approaches to estimating the number of local optima. As we discuss each of these approaches, we will

first describe the approach and present the relevant literature. Then we will discuss and critique efforts in each area as to their relative strengths and weaknesses, characterizing how these approaches simplify or depart from the P-H Probability Model.

3.1 Previous Work in Optimal-Value Estimation

In the optimal-value estimation literature there are been four general approaches. The three most common approaches use a continuous probability distribution for W, differing in the basis they use for selection of the specific continuous probability distribution. The simplest approach, which we call the truncation-point approach, makes no assumptions about the functional form of the probability distribution except that it is continuous and has a lower truncation point at θ. The next approach using a continuous random variable for W is the extreme-value-theory approach. It relies upon a classic statistical result by Fisher-Tippett (1928) that provides a limiting distribution for the minima of n samples of size m, as m approaches infinity. The third approach, the limiting-distribution approach, selects a probability distribution that arguably should be an appropriate limiting model as the number of heuristic runs increases towards infinity. The final approach to optimal-value estimation uses a discrete multinomial probability model.

The authors we discuss wrote from a variety of perspectives, using different simplifying assumptions and modeling techniques. As we review their work, we categorize them with respect to Figure 2 and the taxonomy of perspectives from Section 1.7. We will also comment on how their modeling approaches fit into the structure of the P-H Probability Model and the discussion in Chapter 2. Moreover, we are often able to state characteristics of $\mathcal{I}$ and $\mathcal{H}$ that affect the amount of error resulting under the simplifications proposed by each author.

3.1.1 The Truncation-Point Approach

The truncation-point approach is an optimal-value estimation approach that begins with the natural choice of the sample minimum as an estimate of the truncation point (theoretical lower bound) of a random variable. Since the sample minimum is expected to overestimate the optimal value, i.e., to be positively biased in finite samples, a series

expansion approach is used to develop a new estimator with the linear bias term removed. (Robson and Whitlock 1964)

Dannenbring (1977) applies the truncation-point approach to optimal-value estimation on a single instance of a combinatorial optimization problem, that is, from the heuristic user's perspective. He also extends the approach by incorporating a single solution value observation from a stronger heuristic that usually gives a better solution. In this case, the underlying experiment is of the form $\mathcal{I}\times\mathcal{H}_{\mathrm{PRS}}{}^{n}\times\mathcal{H}_{\mathrm{S}}$. $\mathcal{H}_{\mathrm{PRS}}{}^{n}$ is n replications of the Pure Random Sampling heuristic (PRS), which selects feasible solutions independently and distributed according to their representation in the population of feasible solutions. $\mathcal{H}_{\mathrm{S}}$ is a single replication of some stronger heuristic procedure. Denoting the kth order statistic of the PRS solution value sample as $w[k, n]$ and the solution value observation from the stronger heuristic as s yields the following expressions for Dannenbring's estimators:

$$\hat{\theta}_{SAMP} = w[1,n]-(w[2,n]-w[1,n])$$

$$\hat{\theta}_{HSAMP} = \left\{\begin{array}{ll} s-(w[1,n]-s), & \text{if } s < w[1,n] \\ w[1,n]-(s-w[1,n]), & \text{if } w[1,n] < s < w[2,n] \\ \hat{\theta}_{SAMP}, & \text{if } w[2,n] < s \end{array}\right\}$$

$$\hat{\theta}_{MINHS} = \min\{\hat{\theta}_{SAMP}, \hat{\theta}_{HSAMP}\}$$

$$\hat{\theta}_{AVG} = \frac{\hat{\theta}_{SAMP} + \hat{\theta}_{MINHS}}{2}$$

Nydick and Weiss (1994) provide analytical results for the bias and MSE of the four truncation-point estimators suggested by Dannenbring (1977), subject to assumptions on the underlying solution value distribution. Their results support the conclusion that $\hat{\theta}_{AVG}$ is the most robust of these estimators, with all of Dannenbring's estimators performing better when s is obtained from a relatively weak heuristic. They suggest that future efforts in the area of solution value estimation focus on developing estimators that account for the bias in the heuristic solution given by s, exploring higher-order versions of Dannenbring's estimators (i.e. incorporating more observations from the strong heuristic), or creating similar versions of other optimal-value estimators (e.g. incorporating an s into the estimators provided by extreme value theory).

3.1.2 Discussion and Critique of the Truncation-Point Approach

The greatest strength of Dannenbring's truncation-point approach is in the desire to incorporate all available information in order to obtain an improved (reduced-bias) point estimate of the optimal value. Unfortunately, without utilizing a probability modeling framework that captures both the behavior of the random sample on W and the single heuristic observation s, there is no means for measuring the uncertainty or confidence associated with any of Dannenbring's estimates. Moreover, since no single truncation-point estimator is dominant across all relationships of s to the sample on W, the practitioner has no means of knowing which estimator to trust and how much trust to place in it. (See Appendix A for a diagram depicting which of Dannenbring's estimators is best under a variety of conditions on s relative to $w[1,n]$ and $w[2,n]$.)

The analytical evaluation provided by Nydick and Weiss (1994) shows clearly that Dannenbring's estimators are most suited for situations when the problem instance solution values have a uniform distribution, and the "strong" heuristic is weak enough to behave as though it were a member of the random sample on W. This is the only situation in which all of Dannenbring's estimators tend towards zero bias and still have very low MSE.

The assumption of a continuous distribution for W under the truncation-point approach is called into question if n approaches ψ and we begin to observe duplicate values in the sample on W. If $s = w[1,n] = w[2,n]$, all of the truncation-point estimators yield $\hat{\theta} = s = w[1,n] = w[2,n]$. This does not seem like an unreasonable point estimate, if s comes from a truly strong heuristic and n is indeed quite large with respect to ψ.

Use of the P-H Probability Model could facilitate estimation techniques using multiple sampling methods on individual problem instances. These distinct sampling methods are distinct heuristics, so they would correspond to different values for the p_i. The locations of $z[i]$ captured in the indicator functions would remain fixed for a given problem instance. One recommendation in Nydick and Weiss (1994) is to "weight" the heuristic solution, s, according to its bias or strength. They would adapt Dannenbring's estimators by multiplying s by a weight inversely proportional to its strength. More appropriately, we could use our prior expectations about the strength of the "strong" heuristic to provide a prior distribution for the p_i vector which applies to that heuristic, while the analogous

vector for the random sampling on W would have a prior reflecting the equal chance of returning any $z[i]$.

3.1.3 The Extreme-Value-Theory Approach

The extreme-value-theory approach is an approach to optimal-value estimation from the heuristic user's perspective. It is based on a classic result by Fisher and Tippett (1928) which was extended by Gumbel (1958). The Fisher-Tippett result deals with taking n independent samples of size m from a continuous population with fixed lower bound θ. If x_i is the minimum observation from sample $i = 1, \ldots, n$, then the asymptotic distribution of x_i (as m increases) is Weibull with location parameter θ, scale parameter β, and shape parameter α. The result requires that $F(X)$ behave like $\beta(X-\theta)^{\alpha}$ as X approaches θ.

Invoking the Fisher-Tippett result, the extreme-value-theory approach estimates the optimal value as the Weibull location parameter. General techniques for parameter estimation such as the method of moments and maximum likelihood may be applied to estimate the Weibull location parameter. Zanakis (1979) provides a simple analytical estimator for the Weibull location parameter that is often used in literature with reports of relative success. The Zanakis estimator is $\hat{\theta}_Z = \dfrac{w[1,n]w[n,n]-\left(w[2,n]\right)^2}{w[1,n]+w[n,n]-2\left(w[2,n]\right)}$.

In its purest form, optimal-value estimation through extreme-value theory has as its experiment $\mathcal{I}\times\mathcal{H}_W{}^n$, where $\mathcal{H}_W{}^n = \mathcal{H}_{\min\{m\,\mathrm{PRS}\}}{}^n$ is such that each of its n replications returns the best solution out of m PRS replications (i.e. $\mathcal{H}_{\min\{m\,\mathrm{PRS}\}}$ returns $\min\{\mathcal{H}_{\mathrm{PRS}}{}^m\}$); however, several authors use this estimation approach when $\mathcal{H}_W{}^n$ is n replications of some local search heuristic from a random initial solution.

Clough (1969) applies the Fisher-Tippett result to interval estimation of the optimal value for individual problem instances with a nonlinear objective and continuous variables. He uses random sampling to provide a sample on W, $\{w(1), w(2), \ldots, w(n)\}$, where $w(i)$ is the best solution value of the i^{th} independent sample of size m, as called for by the Fisher-Tippett result.

McRoberts (1971) applies the result to formulate a point estimate and a stopping rule for local search heuristics applied to combinatorial problem instances (facility layout). Instead of using random sampling alone, McRoberts (1971) uses a two-station interchange heuristic to generate $\{w(1), w(2), \ldots, w(n)\}$, where $w(i)$ is the solution value at the local optima for the i^{th} heuristic run (from an independent random start). Note that the n "samples" used by McRoberts (1971) are likely to be of different sizes rather than all of a constant size m with m large. Moreover, these n "samples" are not truly random samples, since they are completely determined by the starting point and move strategy. McRoberts (1971) recognized the departures from the Fisher-Tippett assumptions and turned to statistical and visual assessment in order to support the validity of the Weibull distribution.

Golden and Alt (1979) integrate the previous work to form confidence intervals on the optimal value, which they evaluate for several instances of the Traveling Salesman Problem (TSP) with known optimal solutions. They form a sample on W using the 2-OPT and Lin-Kernighan local search heuristics (Lawler et al. 1990) as in McRoberts (1971). Golden and Alt (1979) use the data from this $\mathcal{I}\times\mathcal{H}_W{}^n$ experiment to construct approximate $(1 - e^{-n}) \times (100)\%$ confidence intervals of the form $[w[1,n] - b, w[1,n]]$, where b is the estimate of the Weibull scale parameter derived via maximum likelihood. Like McRoberts (1971), Golden and Alt (1979) find that statistical tests do not reject the hypotheses that the set of heuristic solutions is (1) an independent sample and (2) from a Weibull distribution.

Los and Lardinois (1982) extend the confidence interval of Golden and Alt (1979) so that it can be constructed with any desired level of precision. Their interval is of the form $[w[1,n] - b/S, w[1,n]]$, where $S = \left(\frac{n}{\ln(1 - \text{confidence level})}\right)^{1/a}$ and a and b are the maximum likelihood estimates of the Weibull shape and scale parameters.

Los and Lardinois (1982) note that including multiple observations of the same local optima indicates a failure of the assumptions required by Fisher-Tippett (1928). They discuss a distinction between the induced probability distribution and the continuous Weibull "pseudo-distribution" used in the estimation approach, and suggest that including multiple observations of the same local optima may violate independence or the assumption of a continuous distribution. They remedy this in their procedure by using only

the D distinct local optima. They also suggest that the independence assumption required by the Fisher-Tippett result would be met more closely by using a distinct local improvement heuristic for each of the n samples, with the m sample observations provided by the local optima resulting from m random initial feasible solutions.

Interestingly, the empirical work discussed in Los and Lardinois (1982) on transportation network design problems does not apply the idea of using n distinct heuristics in efforts to better satisfy the Fisher-Tippett requirements. In their experiments, a single local search heuristic generates the n samples via n random starts. The observations in each of the n samples are the solution values along the path from the random initial solution to its associated local optima. They use only the $D \leq n$ distinct local optima in estimating the solution value.

By using local search heuristics with varying power (i.e. varied neighborhood size) and comparing the effectiveness of their procedure, Los and Lardinois (1982) find that a heuristic of intermediate power provides the best confidence intervals. A powerful heuristic results in a tighter Weibull distribution and tighter confidence intervals. However, a heuristic that is too powerful often does not provide enough distinct local optima to effectively estimate the Weibull parameters. In contrast, a weak heuristic results in a Weibull with a larger right tail, producing wider confidence intervals. In addition, very weak heuristics are too easily trapped in local optima, resulting in very small values of m which weaken the appeal to the asymptotic Fisher-Tippett result.

Patel and Smith (1983) prove that for continuous-variable constrained minimization problems with Z such that "1) the unique optimal solution $\mathbf{x}^*$ to [the problem instance] is an extreme point of [the k-dimensional polytope imbedded in R^k which defines the feasible region of the problem instance] and 2) [the objective function z] is strictly increasing and continuously differentiable in a neighborhood of $\mathbf{x}^*$", then W has a Weibull asymptotic distribution with shape parameter equal to the dimension of the feasible region and location parameter equal to the optimal value θ, where W has as its heuristic $\mathcal{H}_{\min\{m\,\mathrm{PRS}\}}$, the heuristic procedure that returns the best solution value from a pure random sample of size m on solution vectors.

The Patel and Smith (1983) result shows that the Fisher-Tippett theorem can be appropriately applied to continuous-variable optimization problems when sampling proceeds in subsamples of size m as described in the theorem. It also extends the Fisher-Tippett result in this context by specifying what value the Weibull shape parameter should take in the limiting distribution. Since several authors have commented on the difficulty of estimating the Weibull shape parameter and the significant impact the shape parameter estimate can have on estimates of the location parameter, θ, the Patel and Smith (1983) result is promising. However, the authors point out that the practical significance of their result remains unclear due to notoriously slow convergence to the asymptotic distribution. Moreover, the result applies directly only to continuous-variable optimization problems rather than the combinatorial problems in which we are most interested. (Appendix B presents a geometric justification for the appropriateness of the Weibull model for the left tail of the solution value distribution in continuous-variable optimization that is very much in accordance with Patel and Smith (1983).)

Wilson et al. (2002) apply the extreme-value-theory approach to assessing heuristic quality on 180 instances of flow-line scheduling problems with makespan objectives. One goal of their research is to compare confidence intervals resulting from the analytical estimators of the Weibull parameters proposed by Zanakis (1979) and those resulting when the Weibull parameters are estimated by the method of least squares.

Since Wilson et al. (2002) recognize that the sampling method producing $\{w(1), \ldots, w(n)\}$ affects the confidence intervals, they use three different sampling heuristics. Their basic heuristics are $\mathcal{H}_{\text{PRS}}$ and two local improvement heuristics that begin with PRS then modify the randomly sampled solution through a series of steps that progressively improve its solution value. Their two local improvement heuristics are $\mathcal{H}_{\text{LLI}}$ = a local improvement heuristic that starts with a complete random solution (schedule), and $\mathcal{H}_{\text{BLI}}$ = a local improvement heuristic that explores a random solution only if the lower bound on its ultimate makespan is better than previous bounds. In efforts to best match the assumptions of extreme value theory, they used 500 subsamples of size 500 for the PRS and LLI

heuristics. So the experiments in Wilson et al. (2002) are $\mathcal{I}\times\mathcal{H}_{\min\{500\ \mathrm{PRS}\}}{}^{500}$ and $\mathcal{I}\times\mathcal{H}_{\min\{500\ \mathrm{LLI}\}}{}^{500}$.

Due to the acceptance-rejection stage in the BLI heuristic, Wilson et al. (2002) do not apply the subsampling structure explicitly in this case. Instead, they perform 50,000 iterations with an observation generated only when an initial solution is accepted for further exploration. This experiment is of the form $\mathcal{I}\times\mathcal{H}_{\mathrm{BLI}}{}^{N}$, where N is a random variable. In their application to 180 problem instances, N ranged from 21 to 1701, with an average of 243.

Before proceeding, Wilson et al. (2002) test all samples for independence using a runs test. Out of the 180 problem instances examined, $\mathcal{I}\times\mathcal{H}_{\min\{500\mathrm{PRS}\}}{}^{500}$ produced 169 samples that pass the runs test, $\mathcal{I}\times\mathcal{H}_{\min\{500\mathrm{LLI}\}}{}^{500}$ produced 166 samples that pass the runs test, and $\mathcal{I}\times\mathcal{H}_{\mathrm{BLI}}{}^{N}$ produced 179 samples that pass the runs test.

Once Weibull parameters are estimated via the Zanakis (1979) estimators or the method of least squares, Wilson et al. (2002) test the Weibull goodness of fit using both the Kolmogorov-Smirnov and Anderson-Darling tests. Confidence intervals are only produced in cases that fail to reject both the assumption of independence and the Weibull fit. When least-squares estimators were used with samples from the PRS and LLI heuristics, only one of the 166 problem instances failed the Weibull goodness-of-fit tests. When the Zanakis (1979) estimators were used with samples from the PRS and LLI heuristics, the Weibull fit was rejected 14 times under the Kolmogorov-Smirnov test and an additional 33 times under the Anderson-Darling test.

Since the sample of observations from the BLI heuristic do not have the explicit sub-sampling structure as do the PRS and LLI samples, Wilson et al. (2002) use a batching strategy before fitting a Weibull model. For a batch of size m, the size of the sample of minima is $n = \min\{\lfloor S/m \rfloor, 500\}$, where S is the total number of observations collected by applying the BLI heuristic to that problem instance. Beginning with $m = 1$, the authors fit the Weibull parameters for increasing values of m until the batched sample passes the tests for independence and both tests for Weibull goodness-of-fit at the 95% level, stopping once

n is decreased to 25. Where the un-batched BLI sample has $S < 25$, Wilson et al. (2002) still attempt a fit for $m = 1$. Each of the 179 BLI samples that passed the runs test also passed the goodness-of-fit tests for some value of m. The confidence intervals produced in this fashion used $m = 1$ in 124 cases and $m = 2$ in 45 cases. The remaining ten confidence intervals had a maximum value of 7 for m.

Wilson, et al. (2002) conclude that their empirical work supports the use of extreme value theory in statistical estimation of the optimal value for combinatorial optimization problems. However, they note that even when a sample passes tests for independence and goodness-of-fit for a Weibull model, the fitted Weibull may not accurately reflect the distribution of minima. They also conclude that both the method used to fit the Weibull parameters and the heuristic used to generate the sample observations have a substantial impact on the confidence interval produced. Moreover, since the Anderson-Darling test is more powerful in detecting lack of fit in the tails of the distribution, Wilson et al. (2002) advocate its use over the Kolmogorov-Smirnov test.

3.1.4 Discussion and Critique of the Extreme-Value-Theory Approach

As suggested by authors such as Los and Lardinois (1982), observing the same optimal value multiple times in a sample on W immediately calls into question the theoretical underpinnings of the extreme-value-theory approach to optimal-value estimation for combinatorial problems. In fact, observing duplicate solution values in the sample on W is merely a symptom of the fundamental flaw in applying the Fisher-Tippett (1928) theorem in this context. In its purest form, the Fisher-Tippett theorem has $\mathcal{H}$ as the heuristic that returns the minimum of m PRS replications, and the Fisher-Tippett result is asymptotic with increasing m. Therefore, when the Fisher-Tippett (1928) theorem applies, the Weibull(θ,α,β) can only be expected to be a good fit as m approaches infinity. However, when considering a sample of solution values from a combinatorial optimization problem instance, the assumption of a continuous random variable becomes patently false as m approaches infinity. In fact, once m has grown sufficiently large, the n observations are very likely to be identically the optimal value θ.

We must further caution authors who rely upon statistical tests for goodness-of-fit and independence to justify forming confidence intervals on the optimal value using a Weibull model. Since we know there are fundamental reasons to suspect that the observations resulting from heuristic solution of combinatorial optimization problem instances are not truly Weibull, goodness-of-fit testing only serves to tell us whether our sampling mechanism (heuristic) and sample size are capable of revealing the underlying lack of fit.

Similarly, it is not appropriate to test for independence in a sample that is independent by construction, then omit from further consideration cases that fail the independence test at a 95% level. Wilson, et al. (2002) test for independence in samples generated by applying heuristics from random initial points. Since these samples are all independent given the particular $\mathbb{I}\times\mathcal{H}$ experiment, 5% of the samples could be expected to fail a 95% test for independence. (When the null hypothesis of an independent sample is true, we expect a 95% failure rate if applying a 95% test.) By omitting the 5% that failed the test for independence, Wilson et al. (2002) may have biased their study so that their results were overly favorable.

However, since many authors have reported good results when estimating the optimal value as the location parameter of the Weibull(θ,α,β), we consider when this approach could provide a satisfactory approximation in spite of the theoretical failure of an appeal to Fisher-Tippett (1928). The simplifying assumption of a continuous probability distribution for W is least likely to be objectionable when n and m are very small relative to ψ and $\mathcal{H}$ is fairly weak and highly randomized, since this situation is very unlikely to result in multiple observations of the same solution value. Moreover, the Weibull(θ,α,β) approximation is most appropriate if the problem instance has solution values that are distributed along the solution value axis in an approximately Weibull fashion. If the heuristic is PRS with $n<<\psi$ and $m<<\psi$, then applying the Weibull(θ,α,β) approximation to a sample of n minima from subsamples of size m is best when the problem instance has its solution values distributed according to a Weibull($\theta,\alpha,\beta\cdot m^{1/\alpha}$) distribution, since the

distribution of minima of m Weibull(t,a,b) random variables is Weibull($t,a,b/(m^{1/a})$). (See Appendix C for the derivation of this result.)

There is some empirical and analytical evidence that a Weibull distribution may indeed be an appropriate problem class model for many optimization problem classes, particularly in the left tail of the distribution. (See Appendix B and Appendix D for some examples.) However, it is easy to imagine heuristics that invalidate the appropriateness of a Weibull model for W, since W reflects not only the problem instance distribution but also the behavior of the heuristic upon it. For example, a heuristic that almost always gets trapped in a locally optimal solution far from θ would likely produce a sample on W that fits a distribution truncated significantly above θ. The heuristic practitioner who opts to use a Weibull model for W must realize that for all of these reasons Fisher-Tippett (1928) does not provide any direct theoretical support for the Weibull as an appropriate model for W.

3.1.5 Limiting-Distribution Approaches

Like the extreme-value-theory approach, limiting-distribution approaches form asymptotic confidence statements on the optimal value. Limiting-distribution approaches utilize probability models to which the distribution of solution values is supposed to converge as the total sample size increases towards infinity. Whereas extreme-value-theory methods for optimal-value estimation on minimization problems always use the Weibull distribution, limiting distribution approaches have a variety of different forms.

Robson and Whitlock (1964) develop a limiting-distribution approach to the upper truncation point. Their approach is based on a series expansion approach to reducing bias in the point estimator of the upper truncation point. When revised to estimate a lower truncation point, their estimate is the same as the truncation-point estimator $\hat{\theta}_{SAMP}$, with approximate 100(1-confidence limit)% confidence interval of the form $w[1,n]-\frac{\text{1-confidence level}}{\text{confidence level}}(w[2,n]-w[1,n]) \le \theta \le w[1,n]$. (Derigs, 1985)

Boender et al. (1982) also develop a limiting-distribution approach for estimating θ in the context of nonlinear minimization problem instances with continuous variables. Their approach results in an approximate 100(1-confidence level)% confidence interval of

the form $w[1,n]-((1-\text{confidence level})^{-2}-1)^{-1}(w[2,n]-w[1,n]) \le \theta \le w[1,n]$. (Derigs, 1985)

Van der Watt (1980) proposes a whole class of confidence intervals for an upper truncation point. When revised to estimate the lower confidence limit, these intervals use the first order statistic and the k+1st order statistic, $w[k+1,n]$. The author gives approximate 100(1-confidence level)% confidence intervals of the form $w[1,n]-(((1-\text{confidence level})^{1/k})^{-2}-1)^{-1}(w[k+1,n]-w[1,n]) \le \theta \le w[1,n]$. (Derigs, 1985)

None of the limiting-distribution approaches discussed here were originally developed for assessing heuristic quality on combinatorial optimization problem instances. However, authors have applied these estimation techniques to optimal-value estimation from the perspective of a heuristic user or heuristic practitioner. Derigs (1985) compares deterministic lower bounds to statistical bounds on the optimal value provided by limiting distribution and extreme value theory in the context of the heuristic user. He uses improving heuristics like 2-OPT to provide the required sample on *W*. Ovacik et al. (2000) provide a direct comparison of the Boender et al. (1982) approach from limiting-distribution theory and the Golden and Alt (1979) approach from extreme-value theory when both approaches used a sample on *W* provided by a long simulated annealing run. Since Ovacik et al. (2000) generate problem instances from a random problem generator and make comparisons of the two estimation techniques at the problem class level, their application is more nearly from the perspective of a heuristic researcher only they compared distinct estimation approaches rather than distinct heuristics.

3.1.6 Comparisons of Limiting-Distribution and Extreme-Value-Theory Approaches

Derigs (1985) evaluates lower confidence limits from the limiting distribution approaches of Robson and Whitlock (1964), Boender et al. (1982), and Van der Watt (1980) along with those from the extreme-value-theory approaches of Golden and Alt (1979) and Los and Lardinois (1982) where Weibull parameters are estimated in a variety of different ways. His goal is to compare these statistical lower limits to deterministic lower bounds in order to assess the practical value of statistical optimal-value estimation.

Derigs (1985) evaluates practical value of statistical optimal-value estimation by comparing both computational requirements and properties such as coverage frequencies and tightness of lower limits. He applies these techniques to twelve traveling salesman problem (TSP) instances and fifteen quadratic assignment problem (QAP) instances from known problem libraries. Ten of the TSP instances and all fifteen of the QAP instances had known optimal solutions at the time. This allows Derigs (1985) to show how often each of the statistical lower limits was in fact higher than the optimal value and hence provided an invalid lower limit. The deterministic lower bounds considered can never yield an invalid lower limit.

Derigs (1985) found that the Golden and Alt (1979) extreme-value-theory approach was the only statistical approach that produced a valid lower limit in all of his trials. However, the Golden and Alt (1979) confidence limit was consistently very wide since its associated confidence limit was essentially 100%. Thus, the deterministic lower bounds available for the TSP instances were always tighter than the Golden and Alt (1979) statistical lower limit. However, the deterministic lower bounds were not as tight for the QAP, which is generally considered more difficult than the TSP. Only on the two smallest QAP instances were the deterministic lower bounds tighter than the statistical lower limit from Golden and Alt (1979).

The Los and Lardinois (1982) extreme-value-theory approach was applied with confidence levels ranging from 99.9% to 90%. Even at the 99.9% level, the Los and Lardinois (1982) approach produced invalid lower limits for nearly a fourth of the problem instances considered. However, the Los and Lardinois (1982) intervals were considerably tighter than the Golden and Alt (1979) intervals, so on both problem classes the deterministic lower bound was usually looser than the Los and Lardinois (1982) lower limit. Derigs (1985) suggests that difficulty in estimating the Weibull shape parameter is the reason the Los and Lardinois (1982) approach produces invalid lower limits with such frequency.

The results for the Robson and Whitlock (1964) approach are much like the Golden and Alt (1979) results. On only one of the TSP instances was the 95% Robson and Whitlock (1964) lower confidence limit tighter than the deterministic lower bound, and in that case the Robson and Whitlock (1964) limit was an invalid lower limit. On the QAP

instances, all but one of the Robson and Whitlock (1964) 95% lower confidence limits were tighter than the deterministic lower bounds, and one of the confidence limits was an invalid lower limit. The Boender et al. (1982) 95% lower confidence limits were tighter than the deterministic lower bounds on four of the twelve TSP instances, but two of these limits were invalid lower limits. On the QAP instances the Boender et al. (1982) 95% lower confidence limits were tighter than the deterministic lower bounds thirteen of fifteen times with only one invalid lower limit.

Derigs (1985) applied the Van der Watt (1980) limiting distribution approach at a 95% confidence level for $k = \{3, 5, 10, 20\}$. Increasing values of k led to increases in both the number of times that the 95% Van der Watt (1980) lower confidence limit was tighter than the deterministic lower bound and the number of invalid lower limits produced. For $k = 3$, the Van der Watt (1980) lower limit was tighter than the deterministic lower bound on half of the twelve TSP instances and all but one of the fifteen QAP instances but produced invalid lower limits on four TSP instances and three QAP instances. For $k = 20$, the Van der Watt (1980) lower limit was tighter than the deterministic lower bound on all but two of the TSP instances and all of the QAP instances but produced invalid lower limits on eight of the twelve TSP instances and ten of the fifteen QAP instances. Derigs (1985) categorizes this as a failure of the limiting distribution approaches and suggests that the continuous distribution models they use are less viable as approximations to the underlying discrete distributions.

Ovacik et al. (2000) applied the extreme value theory approach of Golden and Alt (1979) and the limiting distribution approach of Boender et al. (1982) to single machine scheduling problem instances from a random problem generator. Their goal was to illustrate that the intermediate observations in a long simulated annealing run could be used for statistical optimal-value estimation without requiring significant additional computational effort. They also sought to compare the two optimal-value estimation techniques in this context.

Like Derigs (1985), these authors found that the Golden and Alt (1979) approach rarely produced invalid lower limits, largely because the confidence interval from Golden and Alt (1979) is consistently very large and has a very high confidence level. The 95%

Boender et al. (1982) lower confidence limit was more frequently an invalid lower limit, as the width of these 95% confidence intervals was consistently smaller. They also observed that for both statistical lower limits, the mean and variance (across the problem instances within the problem class) decreased with longer simulated annealing runs. This resulted in the 95% Boender et al. (1982) confidence intervals covering the true optimal value more frequently for longer simulated annealing runs than for shorter ones, since the mean and variance of the statistical lower limit were both smaller for the longer simulated annealing runs.

3.1.7 Discussion and Critique of Limiting-Distribution Approaches

Like the extreme-value-theory approach, limiting-distribution approaches use the simplifying assumption that W is a continuous random variable. Therefore, the results of applying a limiting-distribution approach to optimal-value estimation in combinatorial optimization are best when the sample on W shows no signs of the underlying discreteness. This will occur when $\mathcal{H}$ is relatively weak and highly randomized and $n \ll \psi$. However, the asymptotic nature of the proposed probability distributions suggests that samples with relatively small n should not be expected to produce a good fit.

Since the Robson and Whitlock (1964) confidence level is exact for random variables with an underlying Uniform distribution with lower truncation point at θ, this approach could provide a good approximation when the problem instance at hand has its solution values distributed approximately uniformly on some range [θ, upper limit] and $\mathcal{H}$ is as described above with $n \ll \psi$. Unfortunately, we do not expect many combinatorial problem instances to have this structure.

Comparisons of the limiting-distribution approach and the extreme-value-theory approach in both Derigs (1985) and Ovacik et al. (2000) generally reported more favorable results using the Weibull model suggested by extreme-value theory. This is likely due to the appropriateness of a Weibull model for the continuous problem class seen in the P-H probability modeling framework. (See Patel and Smith (1983) and the discussions in

Appendix B and Appendix D for support of the Weibull model for the underlying continuous random variable.)

3.1.8 The Multinomial Approach

Reiter and Sherman (1965) provide the only discrete approach to optimal-value estimation. The general structure they introduce applies to any heuristic that can be viewed as using a neighborhood and successor function, or move structure, to partition the feasible space of a problem instance into a collection of trees. A tree is the collection of solutions that all lead to the same local optimum under the proposed neighborhood and successor function. As with most of the previously discussed approaches, their basic experimental context has sample space $\mathcal{I}\times\mathcal{H}^n$. Their approach to heuristic-quality assessment from the heuristic user's perspective is through developing a stopping rule based on the tradeoff between potential improvement in the best-observed solution value and the cost of obtaining more observations $\mathbb{H}$ from $\mathcal{H}$.

The probability model they use for the heuristic relies heavily on the partitioning of the feasible set of solutions, $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_\psi\}$, into *trees*. A *tree* is a set of solutions that all lead to the same locally optimal solution under this neighborhood and successor structure, so the solution value observed on the t^{th} application of the heuristic is $w = \min\{z[j] \mid \mathbf{x}_j \in \text{Tree}(\mathbf{x}^{(0,t)})\}$, where $\mathbf{x}^{(0,t)}$ is the feasible solution to $\mathbb{I}$ used to initialize $\mathcal{H}$. Reiter and Sherman (1965) utilize prior information about $\mathbb{I}$ or $\mathcal{I}$ to define a finite set of solution values (e.g. all r integers on the range $[L, U]$) that must contain all of $\{z[1], z[2], \ldots, z[\psi]\}$. They give a Bayes-optimal stopping rule if each of the r possible solution values is viewed as equally likely a priori. The rule is to stop if the expected cost after m observations (an expected value calculation depending upon the successor function defining the heuristic and the probability distribution on feasible solutions used to initialize the heuristic) exceeds the expected gain in additional observations, $V_m = \sum_{i<i_m} \frac{1}{m+r}[R(i_m) - R(i)]$, where R is the return function for reduction in the objective function of our minimization problem and i

indexes the r possible solution values, with i_m corresponding to the best solution value observed in the m observations already collected.

The expected value structure in this stopping rule suggests using a posterior mean over $[L, z[i_m]]$ as a point estimate for θ, with interval estimates formed around the posterior mean according to the posterior probabilities. Using a standard expected value would likely result in an estimate for θ that is not one of the r values we have said are the only feasible values for θ. We can correct this inconsistency by selecting as our point estimate the feasible mass point that is closest in solution value to the expected value of the posterior on $[L, z[i_m]]$, returning both of the surrounding feasible mass points if the expected value falls precisely halfway between them. Using the posterior mean results in a point estimate slightly higher than the simple algebraic mean of the i_m feasible solution values on $[L, z[i_m]]$, since $z[i_m]$ has been observed at least once in the sample on W while none of the other values on $[L, z[i_m]]$ have been observed. The point estimate will approach $z[i_m]$ as the number of times $z[i_m]$ occurred in the sample approaches m and m grows so that it eclipses the r "hypothetical observations" in the prior.

An alternative to the posterior mean would be using the posterior mode and the associated highest probability density intervals for θ. Since the posterior support for θ is $[L, z[i_m]]$, and $z[i_m]$ is the only mass point on this range that has been observed at least once in our sample of size m, the posterior mode for θ is $z[i_m]$. The width of the associated highest probability density intervals will reflect the number of times that $w = z[i_m]$ occurred in the sample in that smaller and smaller intervals result as the number of observations at $z[i_m]$ approaches m and m grows to dominate in the $m + r$ "observations" reflecting the combination of prior knowledge and the current sample on W.

In their empirical example, Reiter and Sherman (1965) utilized an equal probability prior, presumably as an attempt to include little prior information in the estimation procedure. However, this prior has very strong implications as to the behavior of $\mathcal{H}$ and the structure of $\mathbb{I} \in \mathcal{I}$. The rectangular prior distribution says that in previous experience with this problem instance and heuristic, each of the r candidate solution values on $[L, U]$ has been "observed" once in a hypothetical prior "sample" of size r. As a result, the

stopping rule emphasizes only the potential return at each of the yet unobserved solution values, with the expected gain given as the sum total of potential gains divided by the m observations actually observed plus the r hypothetical observations in the prior. So larger r results in a strong prior (i.e. one weighted as though it represented a large body of prior knowledge), which in turn leads us to stop our search earlier under the Bayes-optimal stopping rule given by Reiter and Sherman (1965). When little prior information is available on the heuristic-problem combination, the heuristic practitioner is likely to use something like all integer values on $[L, U]$ to prescribe a value for r. In this case, a larger $[L, U]$ results makes the rectangular prior strong in relation to the sample, leading to earlier stopping under the rules given by Reiter and Sherman (1965).

3.1.9 Discussion and Critique of the Multinomial Approach

The Reiter and Sherman (1965) approach is fully capable of handling a high level of discreteness in our sample on W, which we know could result due to a strong heuristic and/or a sample size that is large with respect to ψ. However, the way Reiter and Sherman (1965) have defined their multinomial categories makes it much more difficult to obtain meaningful prior information than it would be under the P-H Probability Model.

This greater difficulty in formulating an informative prior is due primarily to the confounding of problem and heuristic effects in the Reiter and Sherman (1965) structure. In order to implement an informative prior in this model, the practitioner must be able to meaningfully describe his beliefs about the relative size of the tree rooted at each of the r candidate solution values proposed, knowing that some of these candidate values will not be the root to any of the trees either because (1) none of the feasible solutions to this problem instance has this for its objective function value, or (2) this heuristic is structured so that this solution value is not the value of a locally optimal solution.

Formulating an informative prior in this context requires such detailed knowledge about the problem instance and the way the heuristic operates on it, that it is unlikely to be practicable in cases where we do not in fact already know the true value of θ. Unfortunately, using a non-informative prior does not allow us to leverage what we do know a priori about the structure of our problem class and the nature of our heuristic.

Clearly, there are cases in which this approach provides good posterior estimates of θ with non-informative priors. The trivial example occurs if $z[i_m] = L$. Then the posterior distribution tells us that we are certain that θ is at $z[i_m] = L$.

3.2 Previous Work in Estimating the Number of Local Optima

Since different heuristics produce different sets of local optima when applied to an individual problem instance, some authors estimate the number of local optima resulting under distinct heuristics as a way of comparing heuristics. More often, however, the number of local optima is estimated within a stopping procedure for heuristic sampling in the heuristic user's context. In either context, estimating the number of local optima necessarily incorporates a discrete probability model for the random variable W. Rather than focusing on the lower limit of values for W as in optimal-value estimation, number-of-optima estimation focuses on ψ, the number of distinct solutions to the problem instance, or some $\psi' < \psi$, the number of distinct solutions in some subset of solutions such as the set of locally optimal solutions under $\mathcal{H}$.

The number of local optima could be estimated through the Reiter and Sherman (1965) model by estimating the number of p_i greater than zero. In fact, all of the existing literature on estimating the number of local optima is structurally related to the multinomial model like that in Reiter and Sherman (1965) or like that contained in the P-H Probability Model.

3.2.1 Approaches to Estimating the Number of Local Optima

Boender and Rinnooy Kan (1983) present two Bayes-optimal stopping rules for use when heuristic sampling is applied to a nonlinear optimization problem instance on continuous variables. Their first stopping rule applies to the question of when the practitioner should stop searching for the optimal solution. Their second stopping rule applies to the question of when the practitioner should stop sampling in order to estimate the number of local optima induced on that problem instance by the specified heuristic. Boender and Rinnooy Kan (1983) begin with a multinomial model similar to that in Section 3.1.8. In each case, they use a posterior distribution on the number of local optima

and the relative probability of encountering each of the local optima (ψ and $p_1, \ldots p_\psi$ in our notation). The primary difference between their two decision rules is that the first rule uses a utility function that includes the cost of continued sampling and the cost of stopping with a sub-optimal solution whereas the second rule includes the cost of continued sampling and the cost of stopping with an inaccurate estimate of the number of local optima.

3.2.2 Discussion and Critique of Approaches to Estimating the Number of Local Optima

Clearly, the approaches that estimate the number of local optima simplify the heuristic-quality assessment by ignoring the importance of differences in solution value magnitude. As a result, these approaches cannot effectively be used to assess how much potential error remains when stopping with a heuristic solution that is not proven to be the optimal solution.

Boender and Rinnooy Kan (1983) employ a utility function representing either the cost of stopping before finding the optimal solution or the cost of stopping with the current estimate of the number of local optima. Without incorporating solution value, it is hard to imagine how this utility function could appropriately capture either type of cost. Moreover, if an estimate of the number of local optima were selected as a means of comparing heuristic classes, it is not clear whether a practitioner should prefer a heuristic that tends to result in more or fewer local optima for a given problem class. A heuristic that results in fewer local optima has a greater chance of returning the optimal solution in any single realization. However, a heuristic that results in more local optima will often take less computational time and effort to produce an individual realization. Moreover, number-of-optima estimation makes no statements about the relative quality of local optima returned, except that (at least) one of the local optima is also the global optimum. Therefore number of optima estimation will not reveal when a heuristic that returns a small number of local optima may in fact have a fairly high probability of returning a solution that is very far from the optimal solution.

3.3 Summary of Previous Statistical Approaches to Assessing Heuristic Quality

Although all of the heuristic-quality assessment approaches in previous literature can sometimes provide accurate measures of heuristic quality, they all have their drawbacks. The potential for substantial modeling error and related complications is shown in how each approach differs from the P-H Probability Model for heuristic solution of combinatorial optimization.

Since the truncation-point, extreme-value-theory, and limiting-distribution approaches all simplify the P-H Probability Model by applying a continuous probability model to W, all of these approaches are inappropriate as the heuristic providing our sample becomes very powerful and/or our sample size becomes large relative to the number of discrete solutions. When our heuristic is weak (i.e. similar in nature to PRS) and our sample size is relatively small (i.e. much smaller than ψ), continuous approximations may provide reasonable estimates of the optimal value, provided the continuous probability model selected is a good reflection of the combination of the heuristic at hand and the particular combinatorial problem class or instance of interest. Unfortunately, the heuristic practitioner may not know a priori if this is the situation he faces. In fact, the confounding of problem and heuristic effects in previous approaches makes it exceedingly difficult for a practitioner to assess whether he even believes that the continuous probability model applied to W could be a reasonable approximation.

Since the multinomial-based approaches used in estimating the number of optima and by Reiter and Sherman (1965) for optimal-value estimation are based upon a discrete random variable for W, they are totally appropriate for application to samples where either a powerful heuristic or large sample size has led to observation of repeated solutions. The major drawback to estimating the number of optima is that this approach ignores the importance of solution value in distinguishing which solutions we are most interested in exploring.

The multinomial optimal-value estimation approach proposed by Reiter and Sherman (1965) preserves the emphasis on solution values but retains the confounding of problem and heuristic structure on the distribution of W. This makes it particularly difficult to propose meaningful prior distributions for their Bayesian approach. Since our area of

interest is in the far left tail of the solution value range, and we are very likely to have few observations in this range even if our problem instance contains numerous solutions here, a strong and meaningful prior can be critical in producing accurate and precise Bayesian inferences on parameters of the W distribution.

Clearly, there is room for a new approach. This new estimation procedure would remain in the shaded range of Figure 2 by using a discrete probability model for W and emphasizing optimal-value estimation. By remaining true to the P-H Probability Model, the new approach would confront the practitioner directly with the need for external knowledge about the problem class and heuristic behavior in conducting a statistical assessment of heuristic quality. This approach would make explicit how much information the practitioner has and allow him to honestly assess the strength of the resulting inference. Although detailed knowledge on the problem class and heuristic may be difficult to formulate, it is vital to powerful inference in the context of optimal-value estimation.

Since the optimal value is an extreme value, a sample of solution values is not likely to produce powerful inference about it without the assistance of significant external information. However much the heuristic practitioner might hope for a magic bullet approach that takes the sample data on W and provides honest and meaningful inferences on θ, sample data are by their nature a much better indicator of central properties of the underlying probability distribution. Estimation of a distributional truncation point without the use of external information is an exceedingly precarious proposition. Coles and Powell (1996) discuss the use of Bayesian inference in extreme-value modeling. They concede that it may be difficult to specify prior information about extreme behavior of a process. However, they observe that where reliable prior information is available, it can greatly improve the inference since sample data are likely to be rare in the area of interest.

Trustworthy estimates of θ will only be obtained through an approach that accurately represents the likelihood and prior structure for the appropriate $\mathcal{I}\times\mathcal{H}^n$ experiment, incorporating all available external information into the priors and utilizing all available data. Unlike approaches in previous literature, this sort of approach would be less likely to lull the practitioner into a false sense of security by providing strong statistical conclusions when there is actually a substantial amount of undetected modeling error.

By relying on the P-H Probability modeling framework, distinct knowledge about the problem structure and the heuristic behavior could be used separately to formulate prior distributions on the parameters in the W distribution. When the practitioner has little prior knowledge about his problem and heuristic, the heuristic-quality assessment will reflect this fact by producing weaker statistical statements. Although this approach would require more care in implementation than do the simpler models used by previous authors, careful consideration of prior assumptions would lead the user to a better understanding of the strengths and weaknesses in the resulting interval estimates of the optimal value and their associated confidence limits or probability measures. Chapter 4 develops this sort of optimal-value estimation approach based on the P-H Probability Model.

4 OPTIMAL-VALUE ESTIMATION BASED ON THE P-H PROBABILITY MODEL

4.1 An Overview of Our Bayesian Inference P-H Approach to Optimal-Value Estimation

The P-H Probability Model, $P(W = w) \equiv f_W(w) = \sum_{i \ni w = z[i]} p_i = \sum_{i=1}^{\psi} p_i \cdot \Delta(z[i] = w),$ is naturally hierarchical in nature in that it requires higher-level information from two distinct sources, a problem class model and a heuristic class model, in order to accurately represent how the W random variable results from application of the individual heuristic realization to a particular problem instance. This hierarchical flavor along with the disparate sources of information which may be available on the problem and heuristic leads us to propose a Bayesian inference approach using the P-H Probability Model as the likelihood function in the Bayesian structure. Figure 3 outlines a Bayesian inference process as it applies to optimal-value estimation using the P-H modeling framework.

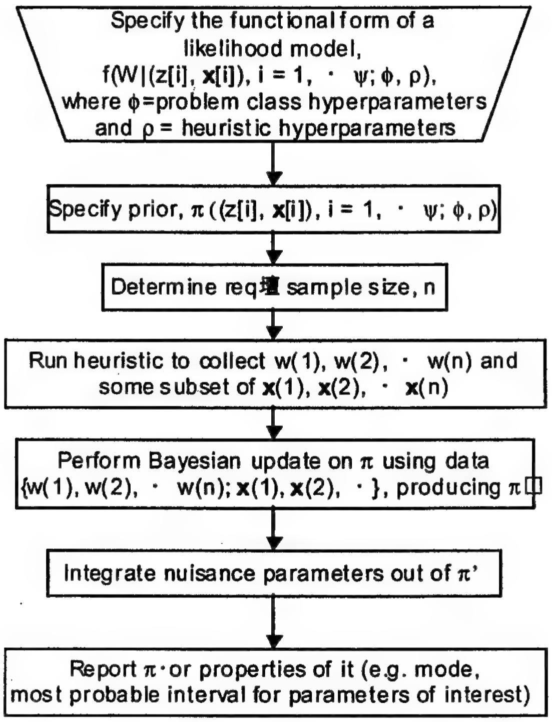

Figure 3. Process Diagram for a Bayesian
Inference Approach to Optimal-Value Estimation
using the P-H Probability Model

The likelihood function for W given the parameters ψ, $z[1],\ldots,z[\psi]$, $p_1,\ldots,p_\psi$ comes directly from the P-H Probability Model as $f_{W|\psi,z[1],\ldots z[\psi],p_1,\ldots,p_\psi}(w)=\sum_{i=1}^{\psi} p_i \cdot \Delta(z[i]=w)$.

Any external information about the problem class becomes a prior distribution on ψ, $z[1]$, …, $z[\psi]$, specifying the hyperparameter vector ϕ, or a probability model for it. Similarly, any external information on the behavior of the heuristic on this problem instance or problem class is used to form a prior on p_1, …, p_ψ by specifying the hyperparameter vector ρ or its probability model. We turn now to ideas on how the problem class and heuristic modeling may be accomplished.

Probability models selected for the problem class and heuristic class priors will be distinctly different if the heuristic practitioner has access to a significant amount of external information than if he has very little information available. The chart in Figure 4 shows the different situations that may arise. Only the extreme cases are labeled in Figure 4, since the issues encountered in these extremes have clear implications for more moderate situations.

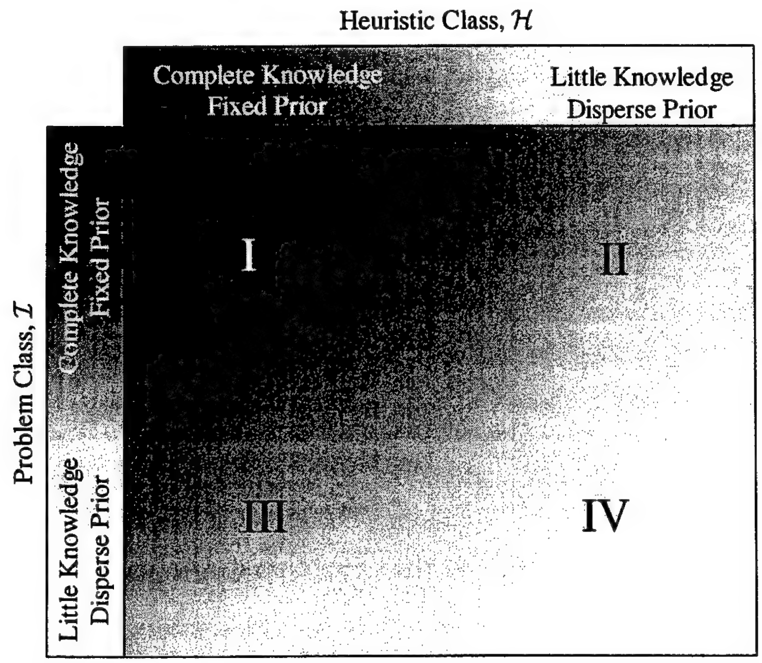

Figure 4. External Information Availability for $\mathcal{I}$ and $\mathcal{H}$ Priors

In Region I, the practitioner has complete knowledge of both the heuristic class and the problem class. At the extreme upper left corner of Region I, this would allow him to specify fixed, constant values for all parameters in the P-H Probability Model. That is, he would have constant values for the location of all feasible solution values, $z[i]$, and for the relative probability of the heuristic returning any particular solution, p_i. The prior distributions are degenerate in this case, and the likelihood model for W stands alone.

In contrast, Region IV of Figure 4 depicts the case where the practitioner has little external information on either the problem class or the heuristic class. In this case, only very disperse prior models can be formulated. These prior models have hyperparameters ϕ and ρ which are varied over a wide range of values in order to reflect the high degree of uncertainty about $z[i]$ and p_i. In the extreme lower left corner of Region IV, only non-informative prior models can be used, with parameter values ranging between $-\infty$ and ∞ for both problem class and heuristic models.

Region II of Figure 4 contains situations where the practitioner understands the problem class structure more fully than the heuristic behavior. These situations result in a

more precise problem class prior model for $z[i]$ and a more disperse model for the heuristic prior on p_i. Similarly, Region III contains situations where the practitioner begins with a better understanding of the heuristic behavior than the problem class structure. These situations result in a more precise specification of the heuristic prior model on p_i values and a disperse form for the problem class prior model on $z[i]$.

Whether the practitioner has complete knowledge, little knowledge, or some intermediate level of external information, the problem class and heuristic class naturally suggest prior models with certain properties or functional forms. We turn now to a discussion of these general forms.

4.2 Models for the Problem Class Prior

The heuristic practitioner may be able to ascertain a fixed constant value for ψ through careful consideration of the problem description and what it means for two solution vectors to be distinct. If he does not have precise knowledge about ψ the practitioner can generally place a lower and upper bound on it and model it as uniform over that range. Hence the prior on ψ varies from a constant value when the practitioner has complete information through a uniform prior between known lower and upper bounds to a uniform prior between $-\infty$ and ∞ when he has no information.

In the upper portion of Region I in Figure 1 the practitioner would be able to specify precisely the ψ discrete $z[i]$ values. Even when little information is available on the precise values of $z[i]$, we know that they are ordered according to their magnitude. Therefore, it is reasonable to view them as order statistics of a sample of size ψ from some underlying problem class distribution. Since the problem geometry suggests a clear relationship among distinct solutions, we would not expect them to behave precisely as the order statistics from an independent sample on the underlying problem class distribution. However, the mathematics of order statistics from an independent sample is very straightforward, so we assume for the time being that this approximation is acceptable.

In many cases, the practitioner can find fixed values for the upper and lower bound on the solution values of feasible solutions to this problem instance, although he may not know much about how tight these bounds are relative to the true range of feasible solution

values. As a result, even in the least informative contexts the practitioner generally will not need to resort to an improper uniform prior on (-∞, ∞) for the problem class model.

When considered together, the above information suggests a prior for ψ, $z[1]$, ..., $z[\psi]$ of the form

$$\begin{aligned}\pi(\psi, z[1],\ldots, z[\psi]) &= \pi(\psi)\pi(z[1],\ldots, z[\psi] \mid \psi)\\ &= \pi(\psi)[\pi(z \mid \phi)]^{\psi}.\end{aligned}$$

where z is any observation from iid sampling on the problem class distribution given by $\pi(z \mid \phi)$. With complete information, this degenerates to a constant for ψ and a constant vector for $z[1]$, ..., $z[\psi]$. When very little information is available, ψ can be modeled as uniform on the set of positive integers with some highly disperse continuous probability distribution used for $\pi(z \mid \phi)$. More often, ψ will be a fixed constant or uniform over a finite range and $\pi(z \mid \phi)$ will be some continuous probability distribution parameterized by the vector ϕ which includes lower bound and upper bound parameters.

In many combinatorial optimization contexts, the problem class model is continuous with the discreteness arising only at the individual instance level. Consider the traveling salesman problem (TSP) for example. Solution values for problem instances in this class can take on any non-negative value on the real line. However, once a particular TSP instance has been selected, solution values for that instance can take on values from some finite collection of discrete values defined by the cost or distance data of that particular instance. Generally, only highly discrete problems require a discrete probability model for solution values at the problem class level.

Since the problem class probability model is generally continuous and we have said the practitioner can supply an upper and lower bound on the feasible solution values, appropriate choices for $\pi(z \mid \phi)$ are bounded continuous probability distributions such as Beta distributions or doubly truncated versions of other continuous distributions. In any case, the vector ϕ represents the hyperparameters of the problem class probability model, reflecting distributional properties such as central tendency, dispersion, and shape. For example, if a doubly truncated Normal distribution were selected as a problem class model,

ϕ would consist of the appropriate mean, standard deviation or variance, lower bound, and upper bound.

4.3 Models for the Heuristic Class Prior

The situation for the heuristic is a bit more complex. We need a joint prior on $p_1,\ldots,p_\psi$ given ψ. We denote the hyperparameters of this prior by the vector ρ. Our likelihood model is like a multinomial model in that it has ψ categories, each of which has probability p_i of being observed. Since the Dirichlet distribution is the conjugate prior for Bayesian inference on a multinomial model, the Dirichlet distribution is an obvious first choice for the form of our prior distribution on $[p_1,\ldots,p_\psi]$. Mathematically, the Dirichlet prior is $\pi(p_1,\ldots,p_\psi \mid \rho=[\gamma_1,\ldots,\gamma_\psi]) = \dfrac{\Gamma\left(\sum_{i=1}^{\psi}\gamma_i\right)}{\prod_{i=1}^{\psi}\Gamma(\gamma_i)}\prod_{i=1}^{\psi}p_i^{\gamma_i-1}$, for $0 \le p_i \forall i$ and $\sum_{i=1}^{\psi}p_i = 1$.

The conjugacy of the Dirichlet prior and the multinomial likelihood means that the posterior distribution of a multinomial model with a Dirichlet prior is another Dirichlet distribution. Specifically, when $\rho = [\gamma_1,\gamma_2,\ldots,\gamma_\psi]$ is the hyperparameter vector for a Dirichlet prior, and the observed data is the vector $\mathbf{v} = [v_1,v_2,\ldots,v_\psi]$, where v_i is the number of times category i appeared in our sample, then the posterior distribution is Dirichlet with parameters $\gamma_1 + v_1,\ \gamma_2 + v_2,\ \ldots,\ \gamma_\psi + v_\psi$.

The $\gamma_1,\gamma_2,\ldots,\gamma_\psi$ are often viewed as the result of some hypothetical sample representing the value and nature of our prior beliefs. That is, the practitioner can determine values for $\gamma_1,\gamma_2,\ldots,\gamma_\psi$ by imagining that his prior knowledge is actually the result of previous sampling with $\gamma_1+\gamma_2+\ldots+\gamma_\psi$ representing the value he attaches to his prior beliefs relative to the sample observations he is about to collect, and the relative size of $\gamma_i / (\gamma_1+\gamma_2+\ldots+\gamma_\psi)$ representing his prior belief as to how likely category i is to be observed. The idea of formulating a Dirichlet prior by imagining a hypothetical prior sample is simple to understand, but perhaps not as simple to do. The heuristic practitioner may not feel confident in any particular choice of $\gamma_1,\gamma_2,\ldots,\gamma_\psi$ without more guidance. At this point, we could look to a non-informative choice of $\gamma_1,\gamma_2,\ldots,\gamma_\psi$ which seeks to "let the data speak

for itself", but it may be more instructive to provide a construct that applies exactly to the question of specifying $p_1,\ldots,p_\psi$ for a particular class of heuristics.

For one class of heuristics that is closely related to random sampling of the feasible solutions, we know how to extend the likelihood for a simple and accurate prior on $p_1,\ldots,p_\psi$. This heuristic class behaves by randomly sampling some η solutions from the set of ψ feasible solutions and returning the single best solution observed. Here we know that

$$p_i = \left(\frac{\psi - i + 1}{\psi}\right)^\eta - \left(\frac{\psi - i}{\psi}\right)^\eta.$$

What are the implications of applying this structure on the set $p_1,\ldots,p_\psi$? Since η represents the size of the internal sample used by the heuristic, it is a positive integer in this simplest version of the structure. When $\eta = 1$, the heuristic is pure random sampling of the feasible solutions, so each solution has $p_i = 1/\psi$. As η increases, the heuristic becomes stronger in that it increasingly favors the better solutions (i.e. those with smaller index). In fact, this structure cannot accommodate a heuristic that has $p_i > p_j$ for some $i > j$.

Since local search heuristics return only some set of $\psi_H \leq \psi$ local optima, and we have no reason to believe that these local optima correspond to the ψ_H smallest solution values, $\{z[1],z[2],\ldots,z[\psi_H]\}$, the $p_i = \left(\frac{\psi - i + 1}{\psi}\right)^\eta - \left(\frac{\psi - i}{\psi}\right)^\eta$ construct cannot be expected to apply directly to these heuristics. However, it may still be useful in specifying the γ_i hyperparameters of a Dirichlet prior as discussed above. When applied in this fashion, the heuristic practitioner views his heuristic as analogous to a heuristic that returns the best solution value from a random sample of size η. Once the heuristic practitioner specifies some range on η which seems to adequately reflect the "strength" of his heuristic, this construct provides a corresponding array of values for the $[p_1,\ldots,p_\psi]$ vector. This range can be translated into a prior on $\rho = [\gamma_1,\ldots,\gamma_\psi]$ quite cleanly after the heuristic practitioner specifies an equivalent prior sample size.

4.4 The Posterior Distribution

As depicted in Figure 3, once the practitioner has formulated an appropriate prior for the problem class and the heuristic he can perform the Bayesian update to obtain a posterior distribution on all the parameters in the model, $\{z[1],\ldots,z[\psi],\psi, \phi, p_1, \ldots,p_\psi, \rho\}$. From first principles, the posterior is given mathematically by

$$\pi'(\psi, z[1],\ldots, z[\psi],\phi, p_1,\ldots, p_\psi, \rho \mid w) =$$
$$\frac{\pi(\psi)\pi(\phi)\pi(z[1],\ldots, z[\psi] \mid \psi,\phi)\pi(\rho)\pi(p_1,\ldots, p_\psi \mid \psi,\rho) f(w \mid \psi, z[1],\ldots, z[\psi], p_1,\ldots, p_\psi)}{\int\cdots\int \pi(\psi)\pi(\phi)\pi(z[1],\ldots, z[\psi] \mid \psi,\phi)\pi(\rho)\pi(p_1,\ldots, p_\psi \mid \psi,\rho) f(w \mid \psi, z[1],\ldots, z[\psi], p_1,\ldots, p_\psi) d\psi \cdots dp_\psi}.$$

Since this is a very complex expression, let's explore various aspects of it in the specific contexts reflected in Figure 5.

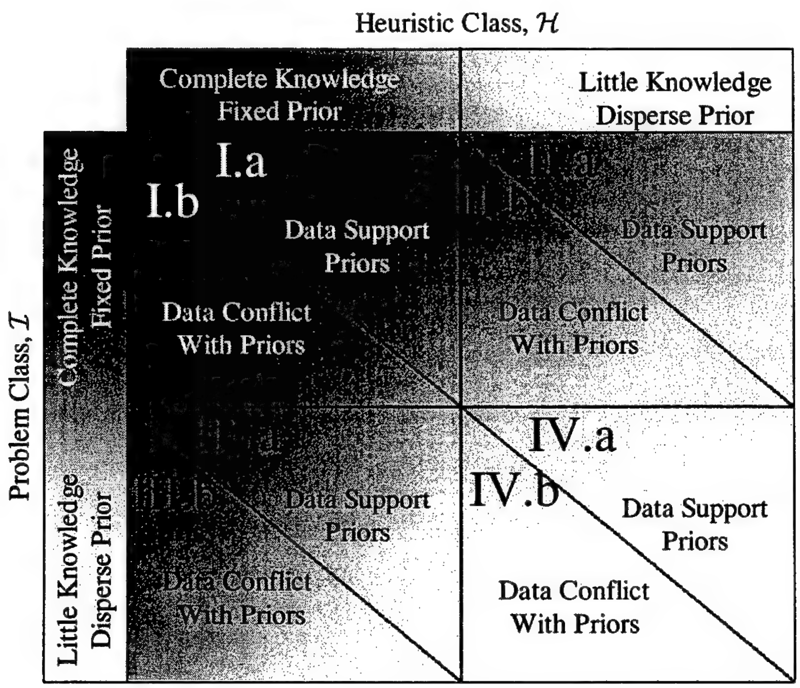

Figure 5. Distinct Situations in Available External Information and Data

In Region I.a of Figure 5, the observed solution-value data supports the precise priors that were specified on both the problem and heuristic. In the extreme upper left corner of Region I, all parameters in the P-H Probability Model are fixed, so observation of

supporting data results in a posterior on the parameters that is identical to the degenerate prior distribution. Elsewhere in Region I, supporting data will produce a posterior that differs from the prior distribution. The result is a joint posterior distribution that is even tighter than the original joint prior distribution, with modes in essentially the same parts of the parameter space. For example, if our prior had a mode for p_1 of 0.5 and a mode for $z[1]$ of 3.5, data that support this prior would result in a posterior with modes again near $p_1 = 0.5$ and $z[1] = 3.5$ with an even greater portion of the probability clustered tightly around these modes than was the case in the prior.

In Region I.b of Figure 5, the observed solution values are somehow in conflict with the specified joint prior. This can occur in different ways with similar results. An extreme example occurs when the problem instance structure is thought to be known completely so that the fixed vector $(z[1], z[2], \ldots, z[50]) = (10, 20, \ldots, 500)$ is the degenerate problem class prior distribution. Then observing $w = 15$ is in direct conflict with the problem class prior. Since a fixed degenerate problem prior was used, this conflict produces a posterior distribution with zero probability everywhere on its allowed parameter space.

A different type of conflict occurs with the same degenerate problem class prior and a heuristic that is thought to always return the optimal solution so that constant values of $p_1 = 1$ and $p_i = 0$ for $i > 1$ are used as the degenerate heuristic class prior distribution. Then observing $w = 20$ is in conflict with the joint prior distribution, indicating that either the problem instance model or the heuristic model is incorrect. Again, with fixed degenerate problem and heuristic prior, the resulting posterior distribution has zero probability everywhere on its allowed parameter space.

In both of these Region I.b examples, if the prior model had been tight but not degenerate, the resulting posterior would be shifted so that the observed data was more likely under the posterior than it had been under the prior. In the first example, when the conflict was clearly a result of error in the problem class prior, the posterior would re-allocate probability over the parameter space so that the probability of some $z[i]$ at $w = 15$ is increased in the posterior. In the second example, the posterior would seek to resolve the conflict by shifting some probability toward areas of the parameter space where at least one

of the following is true: (1) $z[1] = 20$, (2) $p_1 < 1$ and $p_2 > 0$. In this way, the posterior tells the practitioner that either his prior belief about the location of $z[1]$ was wrong, or his prior belief that the heuristic would always return $z[1]$ was wrong, or perhaps both prior beliefs were wrong.

In all cases, the magnitude of the shift in probabilities increases with the amount of conflicting data observed. A single conflicting observation would cause only a slight shift, with a greater change in the posterior resulting from a large conflicting sample. Also, the amount of difference between the prior and posterior decreases with the tightness of the prior distribution. As already mentioned, a degenerate prior with parameters fixed at constant values would not allow any shifting in the posterior. In contrast, a disperse prior would be shifted by even a small sample that conflicted with it. Since all of Region I features relatively tight prior distributions, a substantial amount of conflicting data would be required in order to produce a noticeable shift in posterior probabilities.

In Region IV, the joint prior is very disperse. In the lower right corner, only non-informative prior distributions are available for both the problem class and heuristic class. Since these non-informative prior models provide no information about the parameters of the likelihood, there is no way for data to conflict with them. Throughout Region IV, sample data will have a dramatic effect on the posterior distribution. Small samples could easily result in a misleading posterior distribution, since there is little information in the prior to provide an anchor against "unlucky" samples.

Regions II and III show behavior that is the appropriate blend of behavior in Regions I and IV: conflicting data shifts the posterior away from the prior while supporting data strengthens the prior by tightening it around its mode. As mentioned before, the amount of shift varies directly with the amount of conflicting data observed and inversely with the tightness of the prior distribution.

In Region II, the heuristic portion of the prior is more disperse. As a result, the data will have more of an impact on the posterior along the heuristic dimension in the parameter space, keeping with the greater confidence in the problem class information reflected in the tighter problem class prior. In Region II.b, a conflict like that in the second example would be resolved mostly in favor of the problem class prior, with probability shifted so that the

posterior shows an increased probability that $p_i > 0$ for $i > 1$. In contrast, the data will have more of an impact on the posterior along the problem dimension in the parameter space in Region III, since this region has a more disperse problem class prior reflecting greater confidence in the practitioner's external information about the heuristic. As a result, a conflict like that in the second example would be resolved mostly in favor of the heuristic in Region III.b, shifting probability so that $z[1]$ is more likely to be near 20 in the posterior than in the prior.

4.5 Exploring Our P-H Methodology on a Simplified Example

The best way to build insight on the form of this posterior distribution is by starting with a very simplified version of the situation. By investigating the form of the posterior in the simplified context, we may discover some properties of the posterior that hold true in more complex situations.

We begin with the simplest situation imaginable which roughly corresponds to pure random sampling on a problem instance with only two feasible solutions with solution values on [0,1]. In this case, $\psi = 2$ is fixed and known, the problem class model is Uniform[0,1], and $(p_1, p_2) \sim$ Dirichlet(γ_1=1, γ_2=1), or equivalently, $p_1 \sim$ Uniform[0,1] and $p_2 = 1\text{-}p_1$. What does this joint prior distribution look like? Since ($z[1]$, $z[2]$) is independent of (p_1, p_2), we can plot the prior distribution in two separate three-dimensional plots, the first plot over the ($z[1]$, z[2]) plane and the second plot over the (p_1, p_2) plane.

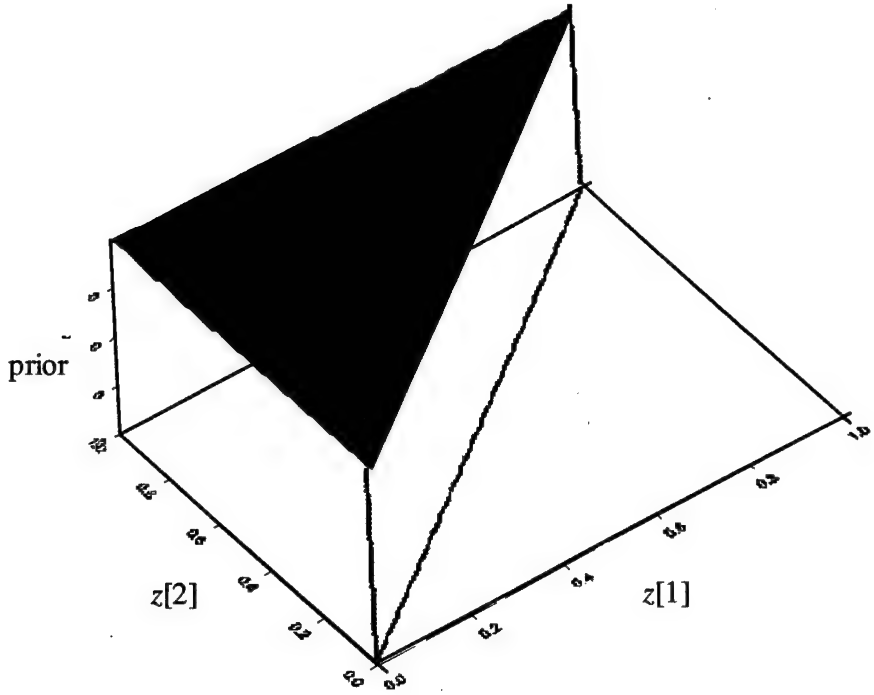

Figure 6. Prior for ($z[1]$, $z[2]$) in the Simplified Example

Figure 6 shows $\pi(z[1], z[2] | \psi{=}2, \phi{=}[LB{=}0, UB{=}1])$. Since this prior views $z[1]$ and $z[2]$ as the full set of order statistics from a sample of $\psi = 2$ Uniform[0,1] random variables, their joint prior is uniform over the region $0 \leq z[1] \leq z[2] \leq 1$.

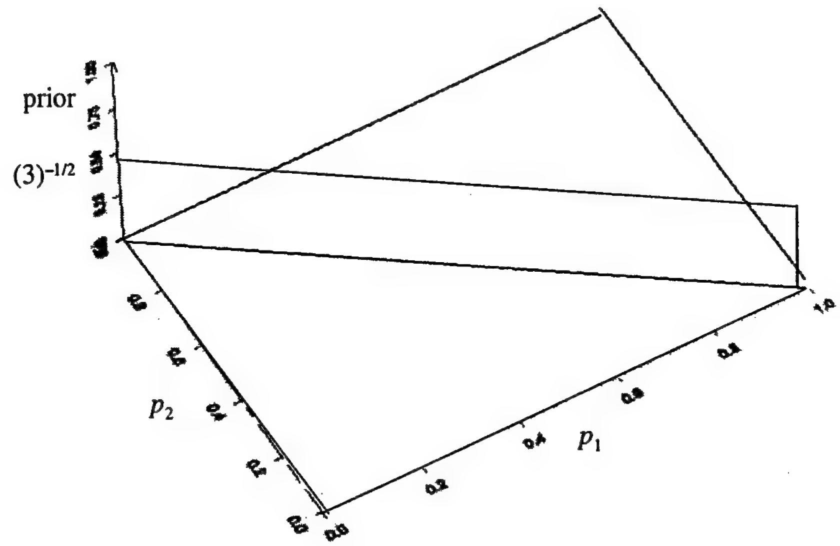

Figure 7. Prior on (p_1, p_2) in the Simplified Example

Figure 7 shows $\pi(p_1, p_2|\psi=2, \rho=[1,1]$ for a Dirichlet distribution). This prior is non-informative in that it gives equal weight to any (p_1, p_2) vector that has $0 \le p_1 \le 1, 0 \le p_2 \le 1$, and $p_1 + p_2 = 1$. However, selecting $\gamma_1 = \gamma_2 = 1$ gives a hypothetical sample size of two, suggesting that our prior knowledge is equivalent in information content to two of the actual observations we are about to collect. For the purposes of a simple illustration, this is acceptable, but practical applications might want to select $\gamma_1 = \gamma_2 = \gamma_\psi = \varepsilon$, where ε is some very small positive real number. Then the hypothetical sample size $\varepsilon\psi$ would reflect much smaller information content for the prior than that of the impending sample.

In order to investigate how this prior is updated into a posterior distribution, suppose we observe data in the form of $w = 0.75$. The resulting joint posterior distribution is given mathematically as

$$\pi'(z[1], z[2], p_1, p_2 \mid \psi = 2, W = 0.75) =$$
$$\frac{\pi(z[1], z[2] \mid \psi = 2)\left[p_1^{\gamma_1} p_2^{\gamma_2 - 1} \cdot \Delta(z[1] = 0.75) + p_1^{\gamma_1 - 1} p_2^{\gamma_2} \cdot \Delta(z[2] = 0.75)\right]}{\iint_{z[1],z[2]} \pi(z[1], z[2] \mid \psi = 2) \iint_{p_1,p_2} \left[p_1^{\gamma_1} p_2^{\gamma_2 - 1} \cdot \Delta(z[1] = 0.75) + p_1^{\gamma_1 - 1} p_2^{\gamma_2} \cdot \Delta(z[2] = 0.75)\right] dp_2 dp_1 dz[2] dz[1]},$$

where the hyperparameters have been suppressed.

Even for this highly simplified example, the posterior is a formidable expression. Let's break it down into more manageable pieces for better understanding. Since our prior on $z[1], z[2]$ given $\psi = 2$ was uniform over the feasible triangle on $[z[1], z[2]]$, let's focus

on interpreting the remaining expression in the numerator,

$p_1^{\gamma_1} p_2^{\gamma_2 - 1} \cdot \Delta(z[1] = 0.75) + p_1^{\gamma_1 - 1} p_2^{\gamma_2} \cdot \Delta(z[2] = 0.75)$. The assumption of continuity for the problem class model implies that $P(Z[1] = Z[2]) = 0$, so only one of the indicator functions in the sum is non-zero. Therefore, we can view this as a statement of cases. In case (i.), $z[1] = 0.75$, so $z[2] > 0.75$, and the posterior Dirichlet parameters are $[\gamma_1 + 1, \gamma_2]$. In case (ii.), $z[2] = 0.75$, so $z[1] < 0.75$, and the posterior Dirichlet parameters are $[\gamma_1, \gamma_2 + 1]$.

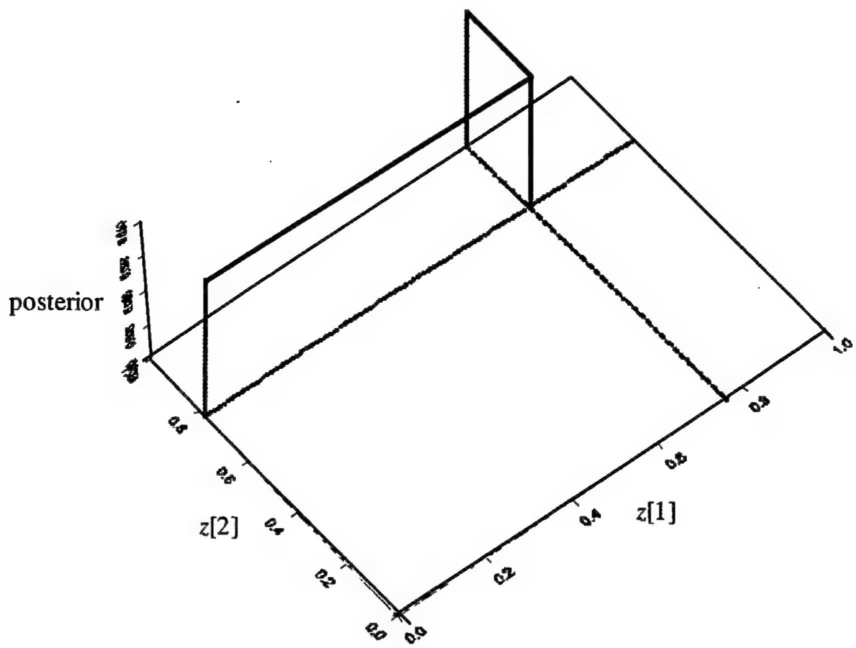

Figure 8. Marginal Posterior on ($z[1]$, $z[2]$) in the Simplified Example

The resulting marginal posterior distribution on ($z[1]$, $z[2]$) is shown in Figure 8. Notice that much of the triangular region depicted in Figure 6 has disappeared. This is because the indicator function in our likelihood model evaluates to 0 everywhere that does not have one of the problem instance solution values equal to our observed w. We have essentially dropped a dimension in going from our prior model on the solution values, $z[1]$, $z[2]$, to the marginal posterior on them. Each of the two "legs" in Figure 8 represents one of the two cases discussed above. Since the problem class prior model was flat, the continuous line segments along the legs are also flat. Since the heuristic prior gave equal

probability to observing either of $z[1]$ or $z[2]$, the line segments on each leg are weighted equally in the posterior distribution so they are at the same height in Figure 8.

Consider how the two-dimensional marginal in Figure 8 would translate into one-dimensional marginal posteriors on the individual $z[1]$ and $z[2]$. The marginal distributions on each of $z[1]$ and $z[2]$ are mixture distributions with some probability at the discrete mass point $z[i] = 0.75$ and the remaining probability allocated to a continuous distribution that is simply a truncated version of the appropriate marginal prior distribution, normalized to integrate to 1.

4.6 Developing the Form of the Posterior Distribution for More General Cases

4.6.1 The Denominator of the Full Posterior

Recall that the full posterior distribution for our P-H Bayesian inference process has the form

$$\pi'(\psi, z[1],\ldots, z[\psi], \phi, p_1,\ldots, p_\psi, \rho \mid w) =$$
$$\frac{\pi(\psi)\pi(\phi)\pi(z[1],\ldots, z[\psi] \mid \psi, \phi)\pi(\rho)\pi(p_1,\ldots, p_\psi \mid \psi, \rho) f(w \mid \psi, z[1],\ldots, z[\psi], p_1,\ldots, p_\psi)}{\int\cdots\int \pi(\psi)\pi(\phi)\pi(z[1],\ldots, z[\psi] \mid \psi, \phi)\pi(\rho)\pi(p_1,\ldots, p_\psi \mid \psi, \rho) f(w \mid \psi, z[1],\ldots, z[\psi], p_1,\ldots, p_\psi) d\psi \cdots dp_\psi}.$$

Put more simply, this equation says

P(*parameters*|*data*) = [P(*parameters*) P(*data*|*parameters*)] / P(*data*). Since we calculate the posterior once data have been observed, the P(*data*) in the denominator is generally a constant, and most Bayesian practitioners omit this and simply normalize the posterior numerator so that it is a proper density function (i.e. integrates to one over the parameter space). However, when our heuristic is pure random sampling, the form of the denominator is fairly straightforward and can be used to check, and help reduce, computational error.

We are treating our problem instance as though it were generated as an iid sample of size ψ from the problem class model. As a result, application of pure random sampling to the solution values of that problem instance produces a sample of size n that may also be viewed as iid according to the problem class model, as long as n is sufficiently smaller than ψ. So P(*data*) is the probability that these particular observations would occur under the

probability distribution used as the problem class prior model, where any hyperparameters are integrated out.

4.6.2 Form of the Marginal Posterior on $\theta = Z[1]$

As shown in Figure 3, the last two steps in a Bayesian inference approach to optimal-value estimation are to eliminate the nuisance parameters and report properties of interest for the remaining marginal posterior distribution. In optimal-value estimation, our primary interest is in $\theta = Z[1]$, so all other parameters are generally viewed as nuisance parameters. Recall from the simplified example with $\psi = 2$, that the marginal posterior distribution for $Z[1]$ is a mixture of a continuous portion where $Z[1] < w[1, n]$ and a discrete spike at $Z[1] = w[1, n]$. Typically, we would attempt to calculate or approximate the full posterior distribution and then derive the marginal posterior distribution for $Z[1]$ by integrating out the other parameters either analytically or numerically. However, since we are modeling $\{Z[1], Z[2], \ldots, Z[\psi]\}$ as order statistics of an iid sample from the problem class probability distribution, we can eliminate the $Z[i]$ for $i > 1$ in a very direct fashion, performing the integration implicitly, then integrate out the remaining parameters.

For the sake of simplicity, we will develop the form of the marginal for the case when ψ and $[p_1, \ldots, p_\psi]$ are known so that $\pi(\psi)$, $\pi(\rho)$, and $\pi(p_1, \ldots, p_\psi | \rho, \psi)$ are not needed in the posterior numerator. We will state the more general form once the development is clear. What remains of the posterior numerator is $\pi(\phi)\, \pi(z[1], \ldots, z[\psi] | \phi, \psi)\, f(w(1), \ldots, w(n) | z[1], \ldots, z[\psi], p_1, \ldots, p_\psi, \phi, \psi) \equiv \pi(\phi)\, P_{\text{Num}}(z[1], \ldots, z[\psi] | \phi, \psi, w(1), \ldots, w(n))$. Since $p_1, \ldots, p_\psi, \psi$ are known constants, we will suppress them in the conditioning. We will also represent $w(1), \ldots, w(n)$ in the conditioning as *data*, for brevity. In the remainder of this section we focus on developing a closed form expression for $P_{\text{Num}}(Z[1] | \phi, \mathit{data})$, revisiting the context of the more general posterior numerator only at the end of the section.

Suppose there are k distinct solution values in our sample of size n. Then denote these k values as $w[1], w[2], \ldots, w[k]$ so that $w[1] < w[2] < \cdots < w[k]$. Let v_j be the number of times $w[j]$ is observed in the sample, so $\sum_{j=1}^{k} v_j = n$. Notice that each of the k observed solution values corresponds to precisely one of $\{Z[1], Z[2], \ldots, Z[\psi]\}$, although we cannot

generally be certain of precisely which $w[j]$ corresponds to what $Z[i]$. Applying the Law of Total Probability, shows that $P_{Num}(Z[1] = w[1]|\phi, data)$ is the sum of $P_{Num}\{(Z[1] = w[1]|\phi, data) \cap$ (this particular assignment of $\{w[1], w[2], ..., w[k]\}$ to $\{Z[1], Z[2], \ldots, Z[\psi]\}\}$ over each possible assignment. For a particular assignment of $\{w[1], w[2], ..., w[k]\}$ to $\{Z[1], Z[2], \ldots, Z[\psi]\}$, $P_{Num}\{(Z[1] = w[1]|\phi, data)$ and (this assignment)$\}= P_{Num}\{(Z[1] = w[1]|\phi, data)$ | (this assignment)$\}$ P_{Num}(this assignment).

Now, $P_{Num}(\text{this assignment}) = \prod_{j=1}^{k} p_i^{v_j} \Delta(Z[i] = w[j] \text{ in this assignment})$, and

$$P_{Num}\{(Z[1] = w[1] \mid \phi, data) \mid (\text{this assignment})\}$$
$$= \prod_{i=1}^{\psi} \left(\begin{array}{l} f_Z(w[j] \mid \phi)\Delta(Z[i] = w[j] \text{ in this assignment}) \\ +(F_Z(w[j+1] \mid \phi) - F_Z(w[j] \mid \phi))\Delta(Z[i] \in (w[j], w[j+1]) \text{ in this assignment}) \\ +(1 - F_Z(w[k] \mid \phi))\Delta(Z[i] > w[k] \text{ in this assignment}) \end{array} \right),$$

where f_Z is the density function for the problem class model and F_Z is the associated CDF. Since only one of the indicator functions reflecting the relationship of $Z[i]$ to $\{w[1], w[2], ..., w[k]\}$ can be true for each $Z[i]$ in any particular assignment, $P_{Num}\{(Z[1] = w[1]|\phi, data)$ | (this assignment)$\}$is a product containing one factor for each $Z[i]$, where each of these factors reflects the probability of that $Z[i]$ relating to $\{w[1], w[2], ..., w[k]\}$ as specified in the assignment.

To help clarify this, suppose we had $\psi = 5$ and $k = 2$, with $v_1 = 1$ and $v_2 = 2$. Then $Z[1] = w[1]$ under the following $\binom{\psi-1}{k-1} = \binom{4}{1} = \frac{4!}{1!3!} = 4$ assignments.

(1.) $\{Z[1] = w[1], Z[2] = w[2], Z[3] > w[2], Z[4] > w[2], Z[5] > w[2]\}$.

(2.) $\{Z[1] = w[1], Z[2] \in (w[1],w[2]), Z[3] = w[2], Z[4] > w[2], Z[5] > w[2]\}$.

(3.) $\{Z[1] = w[1], Z[2] \in (w[1],w[2]), Z[3] \in (w[1],w[2]), Z[4] = w[2], Z[5] > w[2]\}$.

(4.) $\{Z[1] = w[1], Z[2] \in (w[1],w[2]), Z[3] \in (w[1],w[2]), Z[4] \in (w[1],w[2]), Z[5] = w[2]\}$.

The associated probability statements contain factors that evaluate the problem class density for $Z[i] = w[j]$ and factors that are functions of the problem class CDF, reflecting integration over the applicable ranges, for $Z[i] \neq w[j]$. The probability statements for the four assignments in our example are as follows.

(1.) $P_{Num}\{(Z[1]=w[1]|\phi, data) \mid (1.)\} = f_Z(w[1]|\phi)\, f_Z(w[2]|\phi)\, (1-F_Z(w[2]|\phi))^3$.

(2.) $P_{Num}\{(Z[1]=w[1]|\phi, data) \mid (2.)\} = f_Z(w[1]|\phi)\ (F_Z(w[2]|\phi)-F_Z(w[1]|\phi))\ f_Z(w[2]|\phi)\ (1-F_Z(w[2]|\phi))^2$.

(3.) $P_{Num}\{(Z[1]=w[1]|\phi, data) \mid (3.)\} = f_Z(w[1]|\phi)\ (F_Z(w[2]|\phi)-F_Z(w[1]|\phi))^2\ f_Z(w[2]|\phi)\ (1-F_Z(w[2]|\phi))$.

(4.) $P_{Num}\{(Z[1]=w[1]|\phi, data) \mid (4.)\} = f_Z(w[1]|\phi)\ (F_Z(w[2]|\phi)-F_Z(w[1]|\phi))^3\ f_Z(w[2]|\phi)$.

Finally, $P_{Num}(Z[1] = w[1]|\phi, data) = p_1p_2^2\ (f_Z(w[1]|\phi)\ f_Z(w[2]|\phi)\ (1-F_Z(w[2]|\phi))^3)$

$+ p_1p_3^2\ (f_Z(w[1]|\phi)\ (F_Z(w[2]|\phi)-F_Z(w[1]|\phi))\ f_Z(w[2]|\phi)\ (1-F_Z(w[2]|\phi))^2)$

$+ p_1p_4^2\ (f_Z(w[1]|\phi)\ (F_Z(w[2]|\phi)-F_Z(w[1]|\phi))^2\ f_Z(w[2]|\phi)\ (1-F_Z(w[2]|\phi)))$

$+ p_1p_5^2\ (f_Z(w[1]|\phi)\ (F_Z(w[2]|\phi)-F_Z(w[1]|\phi))^3\ f_Z(w[2]|\phi))$.

The continuous portion of the marginal on Z[1] is developed in similar fashion. Let's continue the presentation with our $\psi = 5$, $k = 2$, $v_1 = 1$, $v_2 = 2$ example. There are $\binom{\psi-1}{k} = \binom{4}{2} = \frac{4!}{2!2!} = 6$ assignments for which $Z[1] < w[1]$.

(1.) $\{Z[1] < w[1], Z[2] = w[1], Z[3] = w[2], Z[4] > w[2], Z[5] > w[2]\}$.

(2.) $\{Z[1] < w[1], Z[2] = w[1], Z[3] \in (w[1],w[2]), Z[4] = w[2], Z[5] > w[2]\}$.

(3.) $\{Z[1] < w[1], Z[2] = w[1], Z[3] \in (w[1],w[2]), Z[4] \in (w[1],w[2]), Z[5] = w[2]\}$.

(4.) $\{Z[1] < w[1], Z[2] < w[1], Z[3] = w[1], Z[4] = w[2], Z[5] > w[2]\}$.

(5.) $\{Z[1] < w[1], Z[2] < w[1], Z[3] = w[1], Z[4] \in (w[1],w[2]), Z[5] = w[2]\}$.

(6.) $\{Z[1] < w[1], Z[2] < w[1], Z[3] < w[1], Z[4] = w[1], Z[5] = w[2]\}$.

The associated probability statements are as follows.

(1.) $P_{Num}\{(Z[1]=z<w[1]|\phi, data) \mid (1.)\} = f_Z(z|\phi)\ f_Z(w[1]|\phi)\ f_Z(w[2]|\phi)\ (1-F_Z(w[2]|\phi))^2$.

(2.) $P_{Num}\{(Z[1]=z<w[1]|\phi, data) \mid (2.)\} = f_Z(z|\phi)\ f_Z(w[1]|\phi)\ (F_Z(w[2]|\phi)-F_Z(w[1]|\phi))\ f_Z(w[2]|\phi)\ (1-F_Z(w[2]|\phi))$.

(3.) $P_{Num}\{(Z[1]=z<w[1]|\phi, data) \mid (3.)\} = f_Z(z|\phi)\ f_Z(w[1]|\phi)\ (F_Z(w[2]|\phi)-F_Z(w[1]|\phi))^2\ f_Z(w[2]|\phi)$.

(4.) $P_{Num}\{(Z[1]=z<w[1]|\phi, data) \mid (4.)\} = f_Z(z|\phi)\ (F_Z(w[1]|\phi)-F_Z(z|\phi))\ f_Z(w[1]|\phi)\ f_Z(w[2]|\phi)\ (1-F_Z(w[2]|\phi))$.

(5.) $P_{Num}\{(Z[1]=z<w[1]|\phi, data) \mid (5.)\} = f_Z(z|\phi)\ (F_Z(w[1]|\phi)-F_Z(z|\phi))\ f_Z(w[1]|\phi)\ (F_Z(w[2]|\phi)-F_Z(w[1]|\phi))\ f_Z(w[2]|\phi)$.

(6.) $P_{Num}\{(Z[1]=z<w[1]|\phi, data) \mid (6.)\} = f_Z(z|\phi)\,(F_Z(w[1]|\phi)-F_Z(z|\phi))^2 f_Z(w[1]|\phi)\,f_Z(w[2]|\phi)$.

So, $P_{Num}(Z[1] = z < w[1]|\phi, data) =$

$$\begin{aligned}
&p_2p_3^2\,(f_Z(z|\phi)\,f_Z(w[1]|\phi)\,(F_Z(w[2]|\phi)-F_Z(w[1]|\phi))\,f_Z(w[2]|\phi)\,(1-F_Z(w[2]|\phi)))\\
&+p_2p_4^2\,(f_Z(z|\phi)\,f_Z(w[1]|\phi)\,(F_Z(w[2]|\phi)-F_Z(w[1]|\phi))^2 f_Z(w[2]|\phi))\\
&+p_2p_5^2\,(f_Z(z|\phi)\,f_Z(w[1]|\phi)\,(F_Z(w[2]|\phi)-F_Z(w[1]|\phi))^2 f_Z(w[2]|\phi))\\
&+p_3p_4^2\,(f_Z(z|\phi)\,(F_Z(w[1]|\phi)-F_Z(z|\phi))\,f_Z(w[1]|\phi)\,f_Z(w[2]|\phi)\,(1-F_Z(w[2]|\phi)))\\
&+p_3p_5^2\,(f_Z(z|\phi)\,(F_Z(w[1]|\phi)-F_Z(z|\phi))\,f_Z(w[1]|\phi)\,(F_Z(w[2]|\phi)-F_Z(w[1]|\phi))\,f_Z(w[2]|\phi))\\
&+p_4p_5^2\,(f_Z(z|\phi)\,(F_Z(w[1]|\phi)-F_Z(z|\phi))^2 f_Z(w[1]|\phi)\,f_Z(w[2]|\phi)).
\end{aligned}$$

The general forms for these expressions are

$$P_{Num}(Z[1]=z<w[1]\,|\,\phi,\psi,p_1,\ldots,p_\psi,data)=$$

$$\sum_{i_1=2}^{\psi-k+1}\sum_{i_2=i_1+1}^{\psi-k+2}\sum_{i_3=i_2+1}^{\psi-k+3}\cdots\sum_{i_k=i_{k-1}+1}^{\psi}\left(\left(\begin{array}{l} f_Z(z\,|\,\phi)(F_Z(w[1]\,|\,\phi)-F_Z(z\,|\,\phi))^{i_1-1-1}p_1^{v_1}f_Z(w[1]\,|\,\phi)(1-F_Z(w[k]))^{\psi-i_k}\\ \cdot\prod_{j=2}^{k}p_{i_j}^{v_j}f_Z(w[j]\,|\,\phi)(F_Z(w[j]\,|\,\phi)-F_Z(w[j-1]\,|\,\phi))^{i_j-i_{j-1}-1}\end{array}\right)\right),$$

and

$$P_{Num}(Z[1]=w[1]\,|\,\phi,\psi,p_1,\ldots,p_\psi,data)=$$

$$\sum_{i_2=2}^{\psi-k+2}\sum_{i_3=i_2+1}^{\psi-k+3}\cdots\sum_{i_k=i_{k-1}+1}^{\psi}\left(\left(\begin{array}{l} p_1^{v_1}f_Z(w[1]\,|\,\phi)(1-F_Z(w[k]))^{\psi-i_k}\\ \cdot\prod_{j=2}^{k}p_{i_j}^{v_j}f_Z(w[j]\,|\,\phi)(F_Z(w[j]\,|\,\phi)-F(w[j-1]\,|\,\phi))^{i_j-i_{j-1}-1}\end{array}\right)\right),\text{ with } i_1=1.$$

Note that these general forms remain appropriate when we have non-trivial prior distributions on ψ, ϕ, $p_1, \ldots, p_\psi$, and ρ. So the general form for the posterior numerator with $Z[i]$ marginalized for $i>1$ is

$$\pi(\psi)\pi(\phi)\pi(\rho)\pi(p_1,p_2,\ldots,p_\psi\,|\,\psi,\rho)P_{Num}(Z[1]=z\le w[1]\,|\,\phi,\psi,p_1,\ldots,p_\psi,data),$$

where $P_{Num}(Z[1]=z\le w[1]\,|\,\phi,\psi,p_1,\ldots,p_\psi,data)$

$$=P_{Num}(Z[1]=z<w[1]\,|\,\phi,\psi,p_1,\ldots,p_\psi,data)+P_{Num}(Z[1]=w[1]\,|\,\phi,\psi,p_1,\ldots,p_\psi,data).$$

Although the potentially large number of nested summations make this expression rather cumbersome computationally, the alternative of first forming the full joint posterior then using numerical integration techniques to eliminate $Z[2],\ldots, Z[\psi]$ is even more daunting.

5 APPLYING OUR P-H METHODOLOGY TO A 6-NODE EUCLIDEAN TSP WITH MINISUM OBJECTIVE

In Chapter 4 we developed some insight on the workings of the Bayesian Inference P-H Methodology for optimal-value estimation. In Chapter 5, we apply the methodology to a situation that is a bit more relevant to the heuristic practitioner yet still small enough in scale to provide insight into the process and its results. A brief summary of the results is given in Section 6.1. It is not necessary to read the remainder of Chapter 5 in order to understand the summary in Section 6.1.

In order to apply our methodology in a context that is somewhat familiar to heuristic practitioners yet still simple enough to give clear insight to the behavior of the methodology, we have arbitrarily selected a 6-Node Euclidean TSP instance with minisum objective function from among several we randomly generated. (See Appendix D for more details about the problem generation.) The histogram of solution values for this problem instance is given in Figure 9.

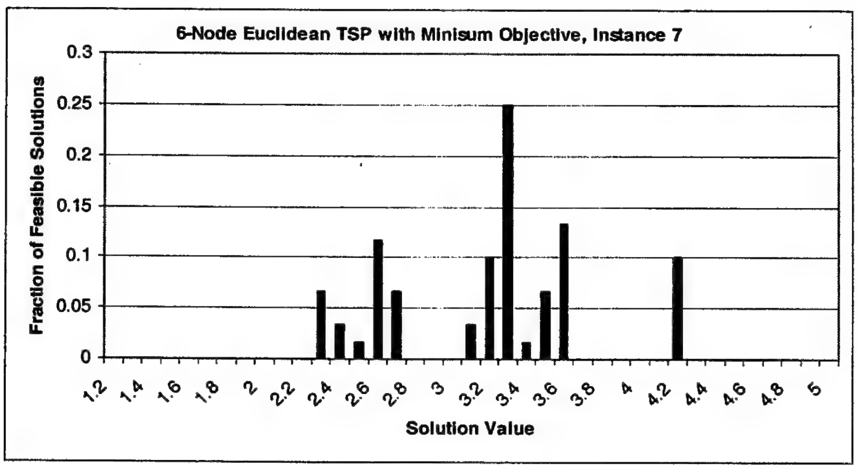

Figure 9. Solution-Value Histogram for Selected Problem Instance

We apply two different heuristic classes to this problem instance: $\mathcal{H}_{\text{PRS}}$ and $\mathcal{H}_{\min\{12\ \text{PRS}\}}$, where $\mathcal{H}_{\min\{12\ \text{PRS}\}}$ is the heuristic that returns a realization that is the best solution value from 12 realizations of $\mathcal{H}_{\text{PRS}}$.

The primary purposes for applying the P-H Methodology to a 6-Node Euclidean TSP instance with minisum objective are to provide more concrete illustration of how prior models can be formulated and to show how the posterior distribution may behave in situations such as those discussed in Section 4.4. As a result, we select our empirical excursions according to the regions depicted in Figure 4 and Figure 5. Recall that Regions III and IV had disperse prior distributions for the problem class while Regions I and II had fixed values for problem class parameters (even $\{z[1], \ldots, z[\psi]\}$ are known constants in extreme portions of Regions I and II). Regions II and IV had disperse prior distributions for the heuristic class while Regions I and II had fixed values for heuristic class parameters.

We would like to provide a clear example of behavior in all of the Regions shown in Figure 5, which incorporates supporting and conflicting data into Regions I, II, III, and IV. The excursions we select attempt to cover all of these combinations. Using Figure 5 to guide our empirical excursions results in an experimental design with at least one excursion in each of Region I.a, I.b, II.a, II.b, III.a, III.b, IV.a, and IV.b, where the "a" regions have data that support the proposed prior distribution while the "b" regions have data that conflict with some aspect of the proposed prior distribution. Due to computing limitations we will discuss more in Chapter 5, we are currently limited to excursions with a very small number of distinct solution values in the sample on W. Most of our empirical excursions use a sample consisting of only two distinct solution values.

The remaining sections in Chapter 4 walk through prior model development for Regions I, II, III, and IV, explain the data used for exploring the "a" and "b" portions of each region, and summarize the insight provided. The presentation is organized according to the regions in Figure 5, sequenced to begin with the extremes of Region I and Region IV then cover Regions II and III.

5.1 Region I Excursions

Recall that Region I contains situations where the practitioner has detailed external knowledge about both the problem class and the heuristic class. At the extreme this allows fixed constants to be specified for all the solution values of the problem instance and the probability that the heuristic returns any of these particular solution values. This extreme is the situation we will explore in our Region I excursions.

As mentioned before, in this extreme upper left corner of Region I, the P-H Probability Model is simply a likelihood function with fixed values for all of its parameters. As a result, there is no updating of a prior distribution on the parameters to a posterior distribution. Rather, presenting the data simply allows us to calculate the probability that the observed solution values actually resulted from the proposed model.

5.1.1 Problem Class Prior Model

To explore the extreme upper left corner of Region I we have enumerated all $\psi =$ 60 solutions to our 6-node Euclidean TSP instance with minisum objective, so we can use the actual solution values $\{z[1], \ldots, z[\psi]\}$ for the problem class prior model. These solution values are {2.20484126, 2.20872027, 2.22373805, 2.22427415, 2.35996567, 2.39441634, 2.49708277, 2.50357152, 2.50980498, 2.51096318, 2.51210055, 2.52017274, 2.52133094, 2.52300441, 2.66492939, 2.67257634, 2.68110536, 2.69085102, 3.08771415, 3.09808191, 3.10431536, 3.11196231, 3.12049134, 3.12216481, 3.13023700, 3.13253257, 3.25671897, 3.25785633, 3.26592852, 3.26876019, 3.27173674, 3.27445755, 3.27499365, 3.27561575, 3.27728922, 3.28148240, 3.28210450, 3.28264060, 3.28377797, 3.28536141, 3.29001143, 3.30153738, 3.40032482, 3.40304562, 3.41859950, 3.41922160, 3.55429102, 3.55544922, 3.56817143, 3.57205044, 3.58706821, 3.58760432, 3.58874169, 3.58989989, 4.16268181, 4.16435529, 4.16497738, 4.18053126, 4.18325207, 4.19713248}

5.1.2 Heuristic Class Prior Models

In order to be able to explore Region I, we needed to use heuristics whose behavior we are capable of knowing explicitly. We selected $\mathcal{H}_{\text{PRS}}$ and $\mathcal{H}_{\min\{12\ \text{PRS}\}}$, because they

both follow the assumptions of $p_i = \left(\frac{\psi - i + 1}{\psi}\right)^{\eta} - \left(\frac{\psi - i}{\psi}\right)^{\eta}$, yet differ substantially in strength.

When our heuristic is $\mathcal{H}_{\text{PRS}}$ and ψ is known exactly, we know $[p_1, \ldots, p_\psi]$ exactly. Pure Random Sampling corresponds to $p_i = \left(\frac{\psi - i + 1}{\psi}\right)^{\eta} - \left(\frac{\psi - i}{\psi}\right)^{\eta}$ with $\eta = 1$, so $p_i = 1/\psi$ for all i. For this problem class, $\psi = 60$ so $p_i = 1/60$ for all i.

Similarly, since ψ is known exactly and $\mathcal{H}_{\min\{12\ \text{PRS}\}}$ returns the best of $\eta = 12$ pure random sampling realizations each time it is applied, we know $[p_1, \ldots, p_\psi]$ exactly. Here $p_i = \left(\frac{60 - i + 1}{60}\right)^{12} - \left(\frac{60 - i}{60}\right)^{12}$, for all i.

5.1.3 Supporting Data Excursions for Region I.a

Excursion 1.) Using the fixed heuristic class prior for $\mathcal{H}_{\text{PRS}}$, we observe the data $w[1] = 3.29001143$, $v_1 = 1$; $w[2] = 3.58874169$, $v_2 = 1$, which are actual observations from two realizations of $\mathcal{H}_{\text{PRS}}$. The likelihood of this sample under the P-H Probability Model with the parameter values given is approximately 0.000278. In fact, this is the likelihood for any two particular solution value observations under the fixed heuristic class prior model for $\mathcal{H}_{\text{PRS}}$, and the fixed problem class prior model, provided that the observed values are among the feasible solution values given in the problem class model.

Excursion 2.) Using the fixed heuristic class prior for $\mathcal{H}_{\min\{12\ \text{PRS}\}}$, observe the data $w[1] = 2.22373805$, $v_1 = 1$; $w[2] = 2.39441634$, $v_2 = 1$, which are actual observations from two realizations of $\mathcal{H}_{\min\{12\ \text{PRS}\}}$. The likelihood of this sample under the P-H Probability Model with the parameter values given is about 0.00872.

5.1.4 Conflicting Data Excursions for Region I.b

Excursion 1.) Using the fixed heuristic class prior for $\mathcal{H}_{\mathrm{PRS}}$, observe the data $w[1] = 2.22373805$, $v_1 = 1$; $w[2] = 2.39441634$, $v_2 = 1$, which are actual observations from two realizations of $\mathcal{H}_{\min\{12\ \mathrm{PRS}\}}$. As in the first supporting data excursion, the likelihood of this sample under the P-H Probability Model with the parameter values given is approximately 0.000278.

Excursion 2.) Using the fixed heuristic class prior for $\mathcal{H}_{\min\{12\ \mathrm{PRS}\}}$, observe the data $w[1] = 3.29001143$, $v_1 = 1$; $w[2] = 3.58874169$, $v_2 = 1$, which are actual observations from two realizations of $\mathcal{H}_{\mathrm{PRS}}$. Under the P-H Probability Model with the parameter values given, this sample is highly unlikely, with a probability of only about 2.18×10^{-17}.

Excursion 3.) Using the fixed heuristic class prior for $\mathcal{H}_{\mathrm{PRS}}$, observe the data $w[1] = 2.5$, $v_1 = 1$; $w[2] = 3.5$, $v_2 = 1$, which are not feasible solution values under our problem class prior model. Since neither of these solution values is in the vector of $z[i]$ parameters, the indicator function in the P-H Probability Model evaluates to zero causing a zero likelihood of this sample under the P-H Probability Model with the given parameter values.

Excursion 4.) Using the fixed heuristic class prior for $\mathcal{H}_{\min\{12\ \mathrm{PRS}\}}$, observe the data $w[1] = 4.18325207 = z[59]$, $v_1 = 1$; $w[2] = 4.19713248 = z[60]$, $v_2 = 1$. Under the P-H Probability Model with the parameter values given, this sample is highly unlikely, with a probability of only about 8.64×10^{-40}.

5.1.5 Summary of Region I Results

As we would hope, likelihood values were considerably smaller for samples with conflicting data that for samples with supporting data. One exception is Excursion 1 in Region I.b. Since the heuristic prior model for this excursion is the fixed model for PRS, there is an equal likelihood of observing any pair of solution value observations from the feasible values in the $z[i]$ vector listed in Section 4.7.1.1. This prior assumption of a very weak heuristic does not allow us to distinguish when a sample actually came from a much stronger heuristic.

5.2 Region IV Excursions

Recall that Region IV contains situations where the practitioner has little external knowledge about both the problem class and the heuristic class. The lack of external knowledge leads to the use of disperse prior distributions on both problem class and heuristic class. In our Region IV excursions we use relatively disperse priors on problem class and heuristic class rather than totally non-informative priors, since this approach illustrates the behavior discussed in Section 4.4 while providing a better illustration of what a practitioner might do in a realistic application of the P-H Methodology.

5.2.1 Problem Class Prior Models

In formulating a relatively disperse problem class prior model for Region IV, we consider the structure of the problem class. For the 6-node Euclidean TSP with minisum objective, the value of ψ is known explicitly. Since selection of the initial node and reverse traversal of the tour do not effect the sum of the arc distances in a tour, a 6-node Euclidean TSP with minisum objective has exactly $\psi = (6-1)!/2 = 60$ distinct solutions. Since ψ is known explicitly, it is treated as a fixed value that contributes to the Bayesian update but is not updated as are parameters which we do not know explicitly.

As mentioned in Section 4.1, a doubly truncated continuous probability distribution makes sense as a problem class prior distribution in many combinatorial optimization contexts. Certainly, since Euclidean TSPs have arc distances that are positive real numbers and the minisum objective function is real-valued and positive as a sum of these arc distances, a doubly truncated continuous probability distribution is appropriate here. Based upon the work of Patel and Smith (1983) and the evidence presented in Appendix B and Appendix D, a three-parameter Weibull distribution truncated above by an upper bound parameter seems like a logical choice. However, this probability model is not the most familiar to heuristic practitioners, so we have chosen instead to use a doubly truncated Normal model for the problem class prior distribution. With this model we need only specify hyperparameters for the upper and lower bounds, the mean, and the standard deviation, i.e. (LB, UB, μ, σ).

We can use concrete physical knowledge of the problem class to specify *LB* and *UB*. Since the locations of nodes in these Euclidean TSPs are determined by drawing independent Uniform[0,1] random variables for the horizontal and vertical coordinates, the physical lower limit on tour distance is $LB = 0$. The largest possible tour distance for a 6-node problem occurs when four of the nodes are at the corners of the $[0,1]\times[0,1]$ square, one node is centered at (0.5, 0.5), and the remaining node is at the midpoint of one side of the square, (0.5, 0) for example. Traversing this tour clockwise from (0,0) gives $UB = 1+1+1+0.5+0.5+\sqrt{2}/2 \approx 4.71$.

Since a practitioner would not know a priori how tight or loose these bounds might be for an individual problem instance, we will not specify (μ, σ) exactly and treat them as fixed in our Bayesian update. Rather, we want to specify prior distributions on these hyperparameters to reflect a prior belief about the problem class structure. We will consider four different models for the hyperparameter prior distributions:

(1.) (μ, σ)~Uniform[$\mu L=LB$, $\mu U=UB$]×Uniform[σL=0.1, σU=15.1],

(2.) (μ, σ)~Uniform[$\mu L=LB$, $\mu U=UB$]×(K/σ on [σL=0.1, σU=15.1]),

(3.) (μ, σ)~

$$\left(K\left(\frac{(UB-\mu)}{(UB-LB)}\right)\left(1-\frac{(UB-\mu)}{(UB-LB)}\right) \text{ on } [\mu L = LB, \mu U = UB]\right)\times\text{Uniform}[\sigma L=0.1,\ \sigma U=15.1],$$

(4.) $(\mu, \sigma)\sim\left(K\left(\frac{(UB-\mu)}{(UB-LB)}\right)\left(1-\frac{(UB-\mu)}{(UB-LB)}\right) \text{ on } [\mu L = LB, \mu U = UB]\right)\times$($K/\sigma$ on [σL=0.1, σU=15.1]), where *K* represents proportionality constants that will not be included explicitly, but become part of the final normalization process.

Prior model (1.) is a naïve disperse prior that specifies fixed upper and lower bounds for both μ and σ and treats all values on the range [μL, μU]×[σL, σU] as equally likely. Since *LB* and *UB* are fixed deterministic limits on solution values for this type of problem, the problem class mean, μ, must also fall between *LB* and *UB*. We know that $\sigma > 0$, but it is difficult to specify an upper limit. Using Excel to calculate and plot the Normal (μ, σ) histogram over [*LB*, *UB*] for various values of (μ, σ) shows us that the probability varies by no more than about 1% on [*LB*, *UB*] for $(\mu, \sigma) > 15$.

Prior model (2.) comes from the transformation invariant non-informative prior on (μ, σ) which is an improper prior proportional to $1/\sigma$ over (-∞, ∞). In the simple case of a one-stage Normal model, the unbounded range of the prior is not a problem because we have a closed form for the resulting posterior distribution. However, in our much more complicated situation, we will need to use numerical methods to calculate the posterior distribution, so we need practical limits on both μ and σ. In this prior we use the same bounds developed for prior model (1.).

Prior model (3.) comes from realizing that our problem instance is very unlikely to have all of its arc distances of $UB = 0$ or $LB = 4.71$, but much more likely to have arc distances near the middle of the [UB, LB] range. This understanding leads to a prior model for μ that is proportional to the product of its distance from UB and its distance from LB.

Prior model (4.) combines the prior on σ from prior model (2.) with the prior on μ from prior model (3.).

5.2.2 Heuristic Class Prior Model

A highly disperse heuristic class prior model would allow all $\psi = 60$ p_i parameters to take on any value on [0, 1] so long as the condition $\sum_{i=1}^{\psi} p_i = 1$ is satisfied. This could be achieved by using $[p_1, p_2, \ldots, p_\psi]$~Dirichlet($\rho = [\gamma_1, \gamma_2, \ldots, \gamma_\psi]$) model with hyperparameter values specified so that $\gamma_1+\gamma_2+\ldots+\gamma_\psi$ is some small value.

A relatively disperse heuristic class prior model can be formed by varying η in the $p_i = \left(\frac{\psi - i + 1}{\psi}\right)^\eta - \left(\frac{\psi - i}{\psi}\right)^\eta$ model. Using a large range on η allows considerable flexibility in heuristic strength, but it does not allow for heuristics that have $p_i > p_j$ for some $i > j$. Since both of our heuristics fit the $p_i = \left(\frac{\psi - i + 1}{\psi}\right)^\eta - \left(\frac{\psi - i}{\psi}\right)^\eta$ structure, we will use $\eta = \{1, 2, \ldots, 12\}$ to form our heuristic class prior model for Region IV, treating all η values on [1, 12] as equally likely.

5.2.3 Supporting Data Excursions for Region IV.a

Excursion 1.) Observe the data $w[1] = 3.29001143$, $v_1 = 1$; $w[2] = 3.58874169$, $v_2 = 1$, which are actual observations from two realizations of $\mathcal{H}_{PRS}$, i.e. $\eta = 1$. Since we have enumerated all feasible solution values for this problem instance, we know that the problem class model ought to have (μ, σ) near (3.2, 0.6) and $z[1]$ near 2.20484126, if it is accurately reflecting this problem instance. We would hope for the posterior distribution to favor $\eta = 1$, since this is the correct parameter value for $\mathcal{H}_{PRS}$. However, since our sample is very small, and we have used a relatively disperse prior model for both the problem class and heuristic class, there is a real possibility that the posterior will be misleading. Figure 10 and Figure 11 show the posterior distributions for η and $Z[1]$, respectively. Recall that the marginal posterior for $Z[1]$ has a discrete spike at $Z[1] = w[1]$. In Figure 11 and $Z[1]$ plots appearing later in Chapter 5, the probability at the spike $Z[1] = w[1]$ is shown as the rightmost datapoint in the series.

Due to the vast number of (μ,σ) combinations considered, and the relatively few grid points with substantial posterior probability, the posterior plot is not given for these parameters. Under all four prior models, the highest probability grid points are those with μ towards its upper limit and σ near its lower limit (generally at around 0.6), with less than 10% of probability distributed among combinations other than at the four following grid points: $(\mu, \sigma) = \{(4.71, 0.6), (4.56, 0.6), (4.40, 0.6), (3.77, 0.1)\}$.

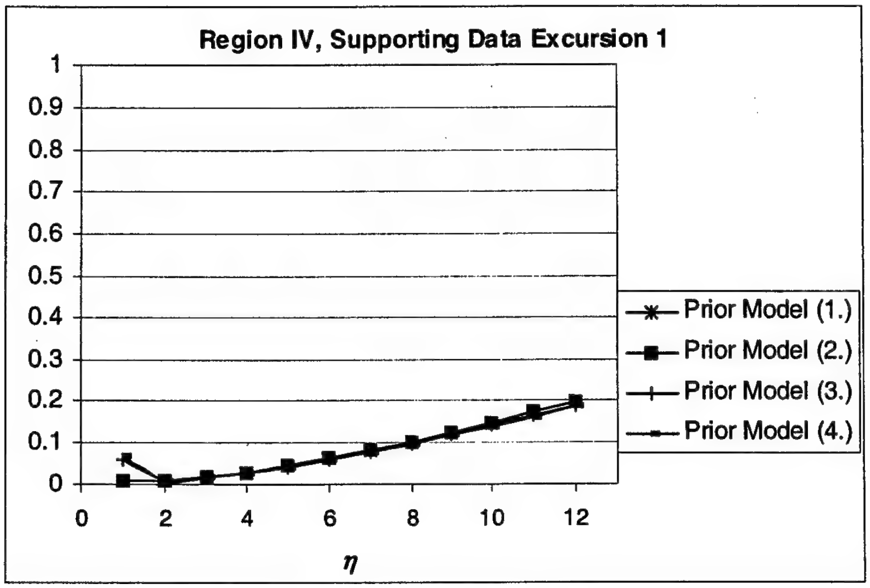

Figure 10. Marginal Posterior Distribution for η from Excursion 1 in Region IV.a

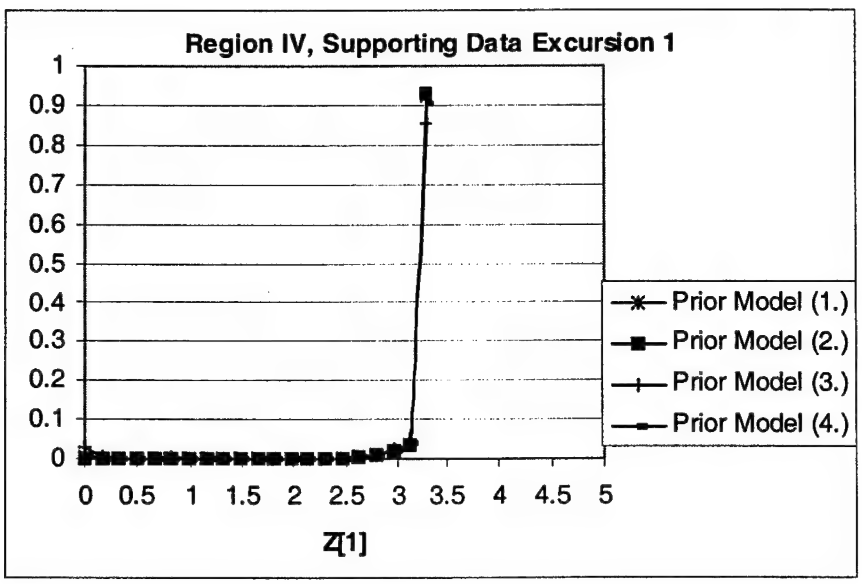

Figure 11. Marginal Posterior Distribution for $Z[1]$ from Excursion 1 in Region IV.a

Note that although the marginal posterior on η favors $\eta = 12$, there is also a noticeable mode at $\eta = 1$ under Prior Model (3.) and Prior Model (4.). More disappointing is the marginal posterior on $Z[1]$, that strongly suggests that $Z[1] = w[1] = 3.29001143$. While this is unfortunate, it should not be terribly surprising considering that the vast majority of our η parameter space featured relatively strong heuristic behavior and our problem class model was sufficiently disperse to contain several problem class parameter combinations that would allow for a strong heuristic providing this sample, strengthening the posterior belief that we observed the optimal value.

Excursion 2.) Observe the data $w[1] = 2.22373805$, $v_1 = 1$; $w[2] = 2.39441634$, $v_2 = 1$, which are actual observations from two realizations of $\mathcal{H}_{\min\{12\text{ PRS}\}}$, i.e. $\eta = 12$. In this excursion, we would hope for the posterior distribution to favor $\eta = 12$, since this is the correct parameter value for $\mathcal{H}_{\min\{12\text{ PRS}\}}$. We would again like for the posterior to favor values of (μ, σ) near (3.2, 0.6), based upon our knowledge of the solution value histogram for the enumeration of this problem instance. This sample is considerably stronger than that presented in Excursion 1, but we still might hope that the marginal posterior distribution for $Z[1]$ favors a value lower than $w[1]$, since the true optimal value is at 2.20484126. However, our sample is still very small and our prior model is disperse, so there is a real possibility that the posterior will again be misleading in some way. Figure 12 and Figure 13 show the posterior distributions for η and $Z[1]$, respectively.

As before, we do not include a posterior plot for (μ,σ) due to the high number of combinations considered and the relatively few grid points that have high posterior probability. The marginal posterior for (μ,σ) is not as tight here as it was in Excursion 1; however, under all four prior models, the highest probability grid point was at (3.46, 0.6). This single grid point held 19-27% of the probability in each posterior. The two next highest probability grid points under Prior Model (1.) and Prior Model (3.) were (3.62, 0.6) and (3.30, 0.6). Under Prior Model (2.) and (4.), (3.62, 0.6) and (2.52, 0.1) were the two next highest probability grid points, due to the effect of the $1/\sigma$ factor in these priors. The three highest probability grid points accounted for a total of 48%, 65%, 81%, and 75% under Prior Model (1.), Prior Model (2.), Prior Model (3.), and Prior Model (4.)

respectively. Note that these grid points have about the same σ value as the highest probability points from Excursion 1, but place μ considerably lower.

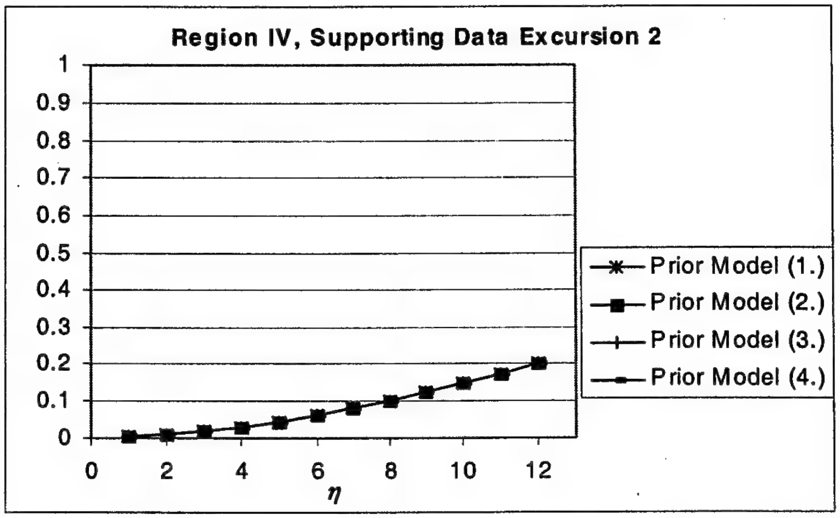

Figure 12. Marginal Posterior for η from Excursion 2 in Region IV.a

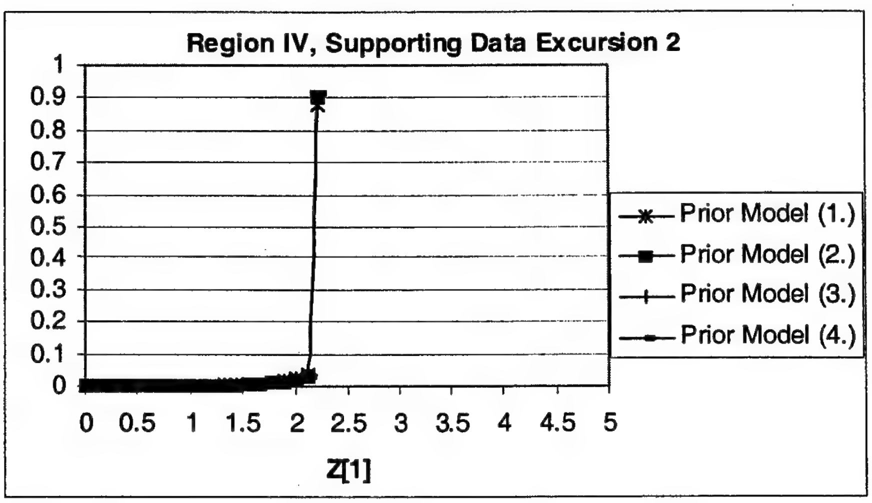

Figure 13. Marginal Posterior for $Z[1]$ from Excursion 2 in Region IV.a

As we hoped, the marginal posterior on η in Figure 12 favors $\eta = 12$ over all other values. The marginal posterior on $Z[1]$, again strongly suggests that $Z[1] = w[1]$, but here $w[1] = 2.22373805$ is much closer to the optimal value. As before, the large number of options for η that indicate a relatively strong heuristic and the dispersion in the problem class prior model led the posterior to favor a strong heuristic and a problem class situated so that the observed data looks likely to have contained the optimal value.

5.2.4 Conflicting Data Excursions for Region IV.b

Excursion 1.) Observe the data $w[1] = 2.20872027$, $v_1 = 1$; $w[2] = 2.22373805$, $v_2 = 1$, which are actual observations from two realizations of $\mathcal{H}_{\min\{25\ \mathrm{PRS}\}}$, i.e. $\eta = 25$. In this excursion, we would hope for the posterior distribution to favor high values of η, since $\eta = 25$ is the correct parameter value for $\mathcal{H}_{\min\{25\ \mathrm{PRS}\}}$. We would again like for the posterior to favor values of (μ, σ) near (3.2, 0.6), based upon our knowledge of the solution value histogram for the enumeration of this problem instance. This sample is so strong that w[1] = $z[2]$ is within 0.2% of the optimal value, so we would be happy with a marginal posterior on $Z[1]$ with much of its probability near $Z[1] = w[1]$. Figure 14 and Figure 15 show the posterior distributions for η and $Z[1]$, respectively.

Again, we omit a posterior plot for (μ,σ) in favor of a brief summary of the posterior mode. The marginal posterior for (μ,σ) is fairly tight here as it was in Excursion 1. Under all four prior models, the highest probability grid point is at (2.36, 0.1). This single grid point holds 43-79% of the probability in each posterior. The two next highest probability grid points under Prior Model (2.), Prior Model (3.), and Prior Model (4.) are (2.52, 0.1) and (3.30, 0.6). Under Prior Model (1.), (3.30, 0.6) and (3.46, 0.6) are the second and third highest probability grid points, followed by (2.52, 0.1). The three highest probability grid points account for a total of 90%, 73%, and 94% of the posterior probability under Prior Model (2.), Prior Model (3.), and Prior Model (4.) respectively. When (2.52, 0.1) is included, the four highest probability grid points account for 69% of the posterior probability under Prior Model (1.). These grid points reveal a posterior

tendency to favor slightly tighter values of σ than we saw in the previous excursions and a somewhat lower value of μ.

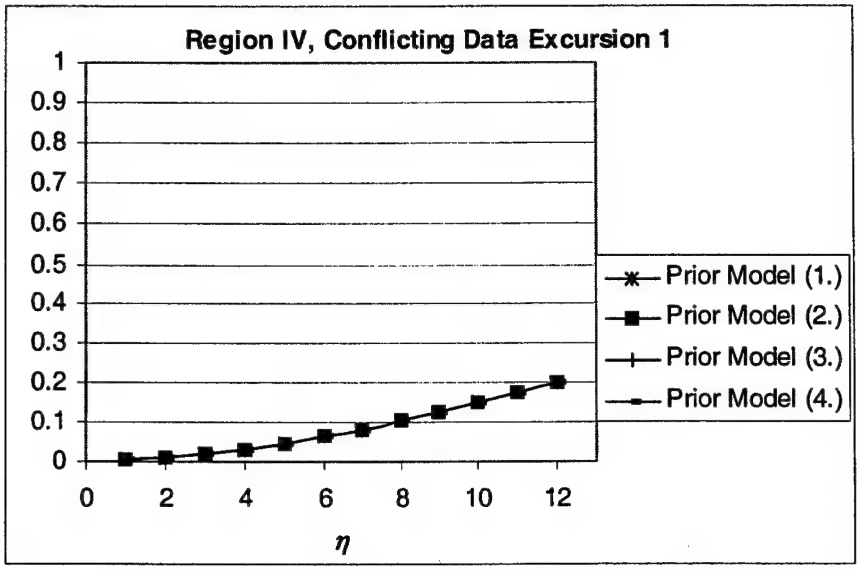

Figure 14. Marginal Posterior Distribution for η from Excursion 1 in Region IV.b

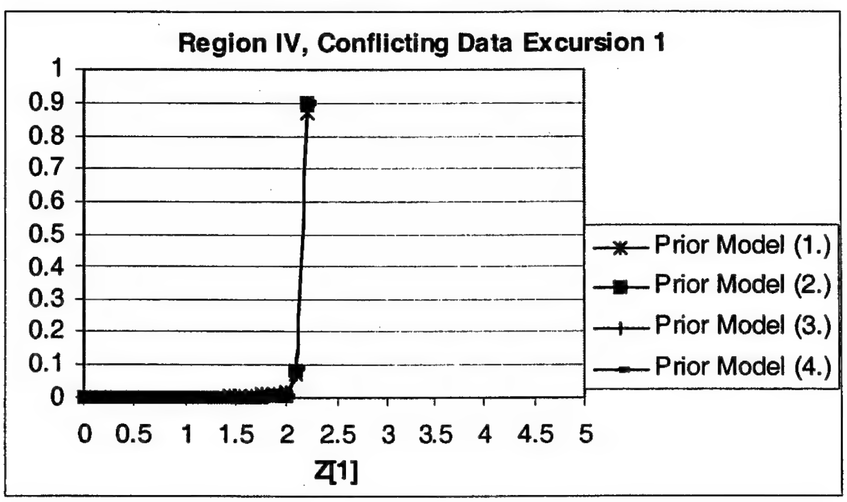

Figure 15. Marginal Posterior Distribution for Z[1] from Excursion 1 in Region IV.b

The results in Figure 14 and Figure 15 are as favorable as we could expect under these disperse prior assumptions. The marginal posterior on η favors the highest value available under the prior model. We know that the sample was actually produced by $\mathcal{H}_{\min\{25\text{ PRS}\}}$, which has a value of η well above the range given in the heuristic prior model. This illustrates that a practitioner should investigate a wider range on any parameter that results in a posterior mode at one of its boundaries, unless the boundary is based on concrete knowledge of the system.

The marginal posterior on $Z[1]$ strongly suggests that the optimal value is at $w[1]$ = 2.20872027. As mentioned above, the optimal value is actually less than 0.2% below $w[1]$. As we saw in the two supporting data excursion, the posterior has favored combinations of problem class parameters that make it very likely that we have observed the optimal value.

Excursion 2.) Observe the data $w[1] = 0 = zL$, $v_1 = 1$; $w[2] = 4.71 = zU$, $v_2 = 1$. Since $w[1] = 0 = zL$, the marginal posterior distribution on $Z[1]$ must allocate all of its probability to $Z[1]= w[1] = 0 = zL$. This could either manifest itself in a posterior that favors a relatively weak heuristic and a problem class with small σ and μ near zero, or a very strong heuristic and a more disperse problem class, perhaps with higher μ. Figure 16 and Figure 17 give the marginal posterior distributions for η and (μ, σ), respectively.

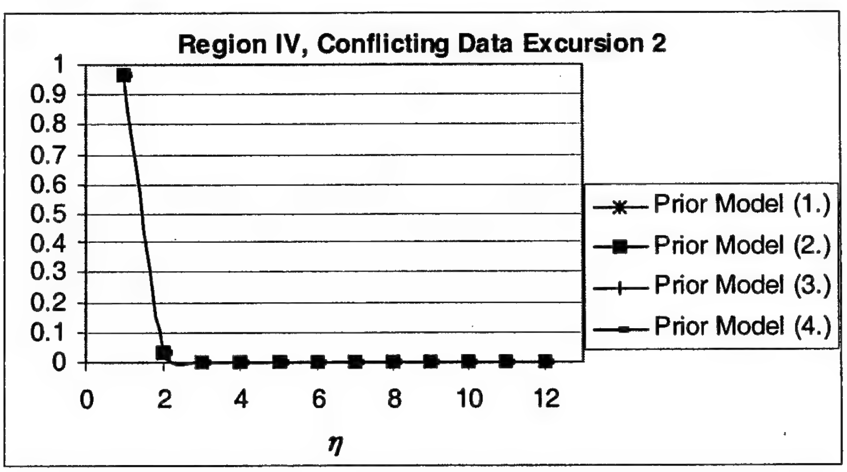

Figure 16. Marginal Posterior Distribution for η from Excursion 2 in Region IV.b

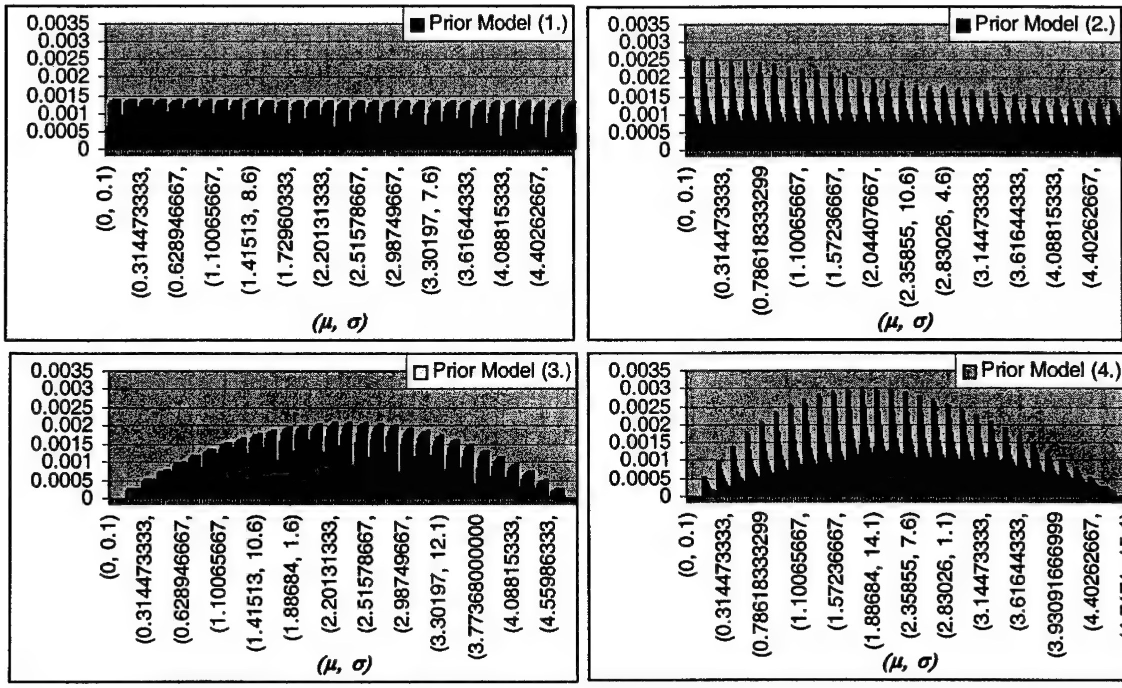

Figure 17. Marginal Posterior Distribution for (μ, σ) from Excursion 2 in Region IV.b

The results in Figure 16 and Figure 17 are a bit perplexing. The posterior distributions favored a weak heuristic, yet they did not indicate a problem class distribution with small σ and low μ. Instead, the posterior distributions on the problem class hyperparameters look very much like their respective prior distributions. This suggests that observing $w[1]$ at the lower bound of the problem class model caused a failure in the updating procedure. If we expect that our problem class model may have a tight lower bound, we must ensure that the prior has non-zero probability at the lower bound, and that our computational approach is capable of handling this sample value.

Excursion 3.) Observe the data $w[1] = 4.18325207 = z[59]$, $v_1 = 1$; $w[2] = 4.19713248 = z[60]$, $v_2 = 1$. Although these observations are actually deep in the upper tail of our problem instance, they are also very close together. As a result, we should not be surprised by a posterior that favors a strong heuristic and a problem class that with a high μ and small σ. We would hope for the marginal posterior on $Z[1]$ to favor values much lower than $w[1] = 4.18325207 = z[59]$, but that will not be the case if the posterior favors a high and tight problem instance.

In fact, the marginal posterior distribution on (μ, σ) has about 99.9% of its probability at (4.40, 0.1) under all four problem class prior models, and the plots in Figure 18 and Figure 19 also indicate a posterior that favors a strong heuristic and a problem instance that is high and tight.

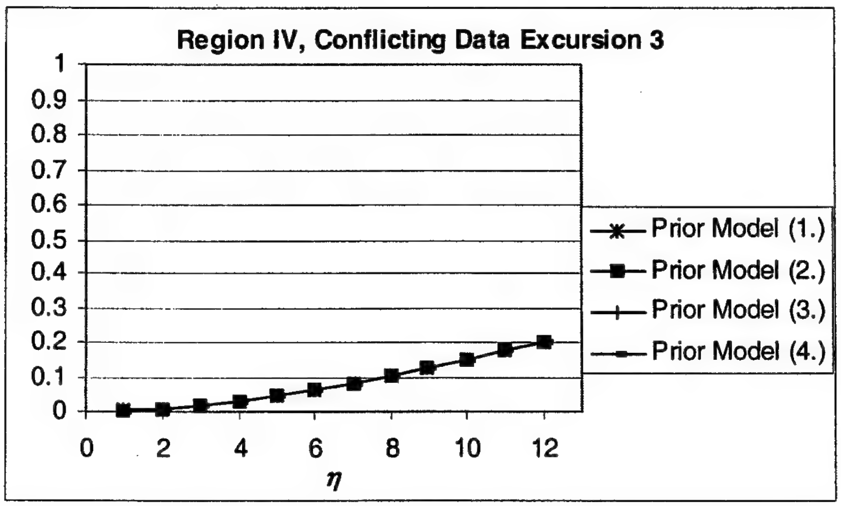

Figure 18. Marginal Posterior Distribution for η
from Excursion 3 in Region IV.b

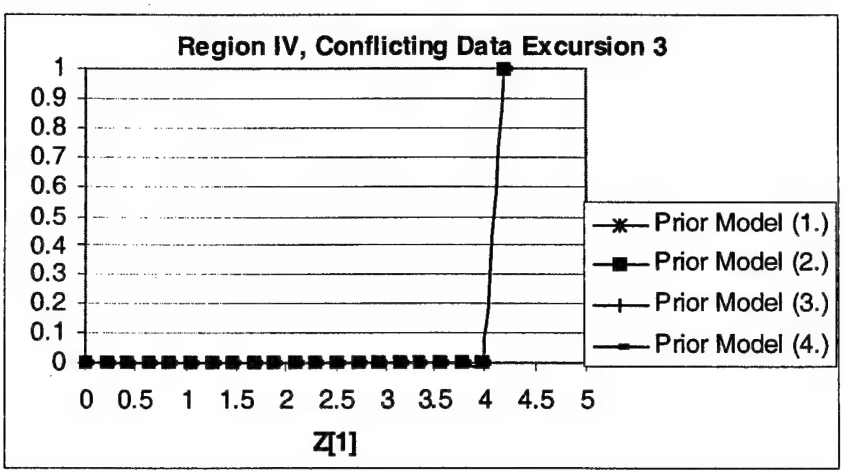

Figure 19. Marginal Posterior Distribution for
Z[1] from Excursion 3 in Region IV.b

Excursion 4.) Observe the data $w[1] = 4.18325207 = z[59]$, $v_1 = 1$; $w[2] = 4.19713248 = z[60]$, $v_2 = 25$. This excursion is fundamentally very similar to Excursion 3, except that it features data that violate the assumption of strictly decreasing values of p_i as i increases. We would like for our posterior distribution to indicate this violation, but it is not clear how that would happen.

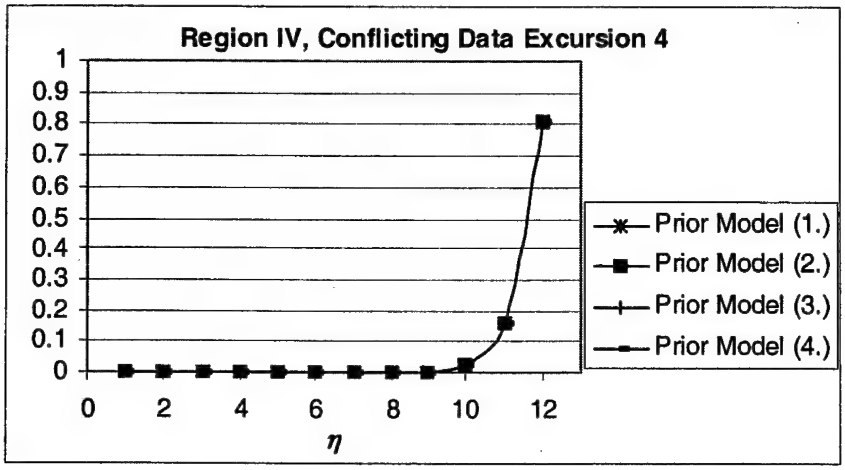

Figure 20. Marginal Posterior Distribution for η from Excursion 4 in Region IV.b

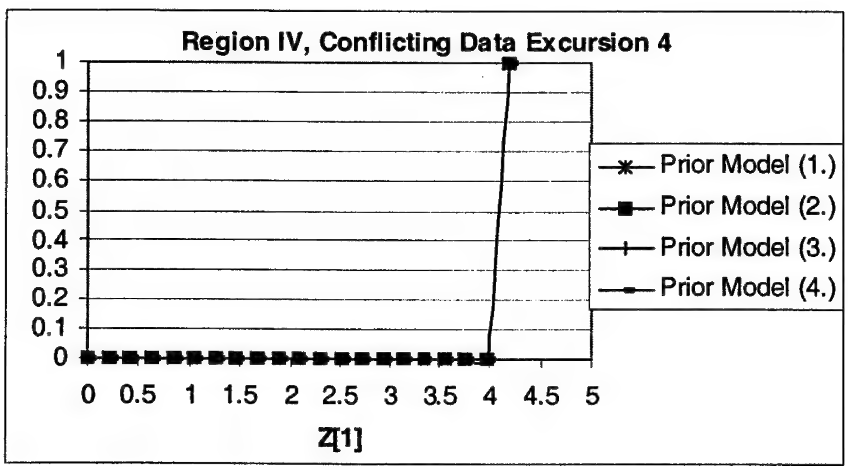

Figure 21. Marginal Posterior Distribution for $Z[1]$ from Excursion 4 in Region IV.b

Instead of a posterior distribution that somehow indicates that our sample violates the underlying structure of the proposed heuristic class prior model, the results in Figure 20 and Figure 21 differ little from the Excursion 3 results in Figure 18 and Figure 19. In this excursion, the posterior favors a strong heuristic even more dramatically, with $\eta = 12$ capturing about 80% of the probability in the marginal posterior on η. As before, the posterior also strongly favors a problem class with $(\mu, \sigma) = (4.40, 0.1)$. This grid point garners over 99.9% of the probability in the marginal posterior distribution on (μ, σ).

5.2.5 Summary of Region IV Results

Many of the Region IV Results show a posterior that misleadingly suggests a strong heuristic and a high probability that we have observed the optimal value. Although we intended for our heuristic class prior model to be relatively non-informative, the functional form for p_i is such that relatively small values of η result in a heuristic that has a strong tendency to return $z[1]$. Therefore, using $\eta = \{1, 2, \ldots, 12\}$ favors a stronger heuristic because most of these equally likely prior values imply a relatively strong heuristic. Moreover, our problem class prior distributions are all disperse enough to include hyperparmeters that define a problem instance distribution that could produce $z[1]$ at or near the $w[1]$ values we presented in our excursions.

We can take two lessons from these results for future applications of this approach. The first lesson is that relatively disperse priors may produce misleading results even if the sample data do not seem particularly "unlucky". The second lesson is that a prior distribution that is intended to be relatively non-informative may in fact contain implicit biases. Recall that this was the case in Reiter and Sherman's (1965) Bayesian application as well.

5.3 Region II Excursions

Region II comprises situations where the practitioner has detailed external knowledge about the problem instance but very little external knowledge about the heuristic class. In our Region II excursions, the problem class prior will be a fixed vector of feasible solution values while the heuristic class prior will be relatively disperse. As a

result, only the heuristic class parameters will have a posterior distribution, with this distribution summarized through the posterior on η.

5.3.1 Problem Class Prior Model

As in Region I, the problem class model for Region II comes from our complete enumeration of the $\psi = 60$ solutions to our 6-node Euclidean TSP instance with minisum objective. We again use the actual solution values $\{z[1], \ldots, z[60]\}$ listed in Section 4.7.1.1 for the problem class prior model.

5.3.2 Heuristic Class Prior Model

We use the same relatively disperse heuristic class prior model in Region II that we used in Region IV. This prior varies η in the $p_i = \left(\frac{\psi - i + 1}{\psi}\right)^\eta - \left(\frac{\psi - i}{\psi}\right)^\eta$ model over $\eta = \{1, 2, \ldots, 12\}$.

5.3.3 Supporting Data Excursions for Region II.a

Excursion 1.) Observe the data $w[1] = 3.29001143$, $v_1 = 1$; $w[2] = 3.58874169$, $v_2 = 1$, which are actual observations from two realizations of $\mathcal{H}_{\mathrm{PRS}}$, i.e. $\eta = 1$. This excursion should have a posterior that favors $\eta = 1$. Figure 22 shows this to be the case.

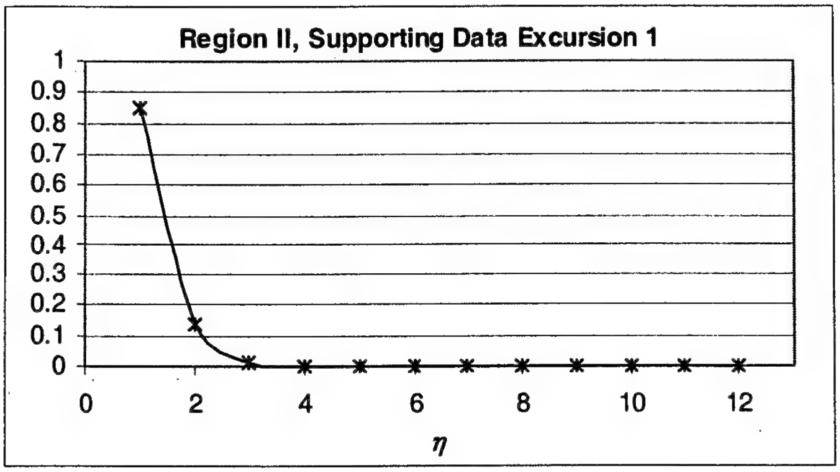

Figure 22. Marginal Posterior Distribution for η
from Excursion 1 in Region II.a

Excursion 2.) Observe the data $w[1] = 2.22373805$, $v_1 = 1$; $w[2] = 2.39441634$, $v_2 = 1$, which are actual observations from two realizations of $\mathcal{H}_{\min\{12\,\mathrm{PRS}\}}$, i.e. $\eta = 12$. This excursion should have a posterior distribution that favors $\eta = 12$. Although the posterior shown in Figure 23 is relatively flat, it does have its mode at $\eta = 12$.

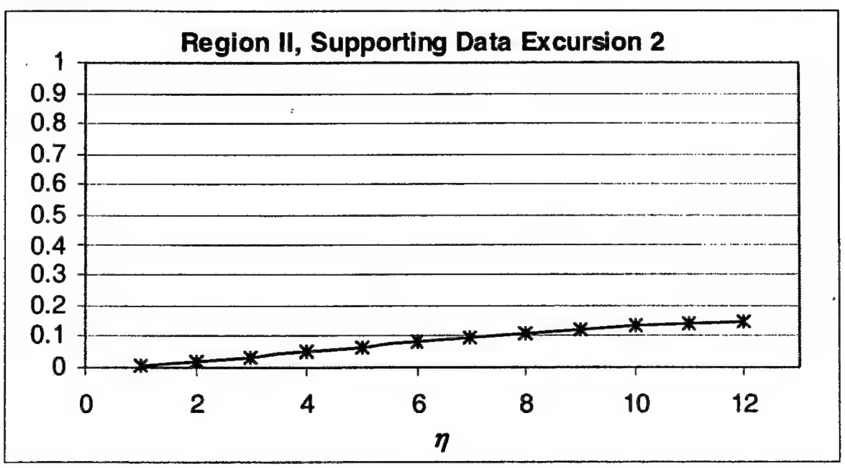

Figure 23. Marginal Posterior Distribution for η
from Excursion 2 in Region II.a

5.3.4 Conflicting Data Excursions for Region II.b

Excursion 1.) Observe the data $w[1] = 2.20872027$, $v_1 = 1$; $w[2] = 2.22373805$, $v_2 = 1$, which are actual observations from two realizations of $\mathcal{H}_{\min\{25\text{ PRS}\}}$, i.e. $\eta = 25$. Since $\eta = 25$ is above the upper limit on η used in the heuristic class prior distribution, we would expect to see the posterior mode at the upper limit of $\eta = 12$. Figure 24 has the appropriate mode, although we might have expected the posterior probability at the mode to be a bit higher.

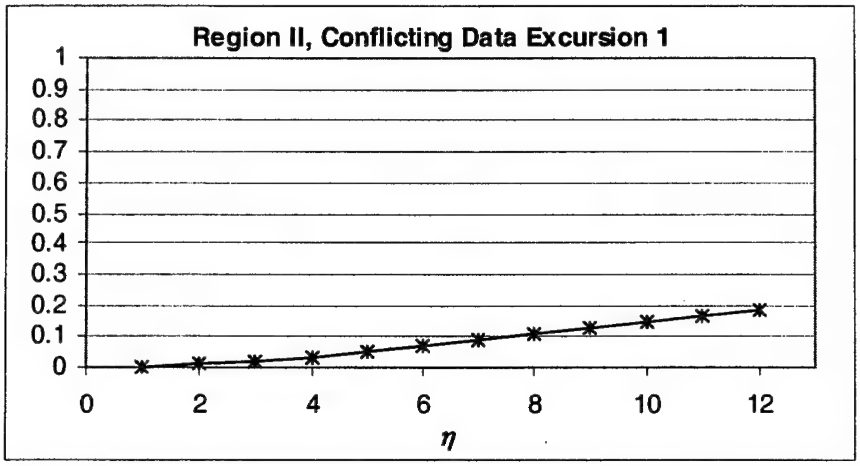

Figure 24. Marginal Posterior Distribution for η from Excursion 1 in Region II.b

Excursion 2.) Observe the data $w[1] = 2.5$, $v_1 = 1$; $w[2] = 3.5$, $v_2 = 1$, which are not feasible solution values under our problem class prior model. Since all of the indicator functions in our likelihood model evaluate to zero for this sample, the posterior distribution must collapse as shown in Figure 25.

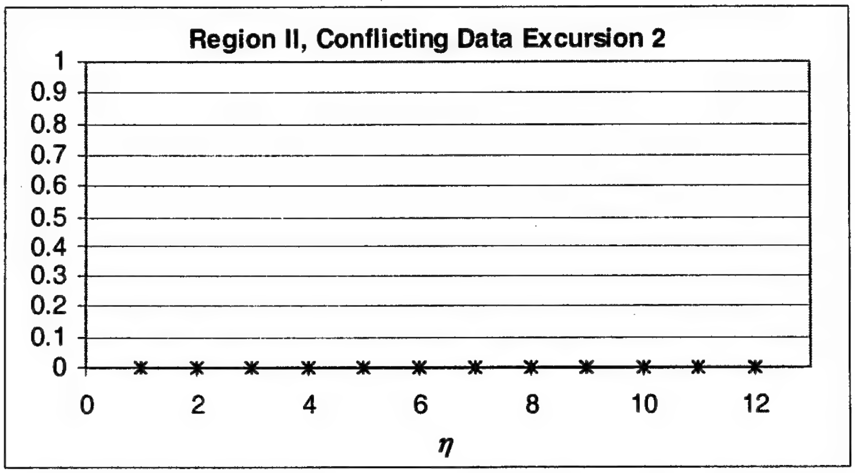

Figure 25. Marginal Posterior Distribution for η
from Excursion 2 in Region II.b

Excursion 3.) Observe the data $w[1] = 4.18325207 = z[59]$, $v_1 = 1$; $w[2] = 4.19713248 = z[60]$, $v_2 = 1$. In Region IV.b Excursion 3, this sample led the posterior to favor a strong heuristic along with a high and tight distribution for problem instance solution values. However, now that the problem instance is known perfectly, the posterior should instead favor a very weak heuristic. Indeed, Figure 26 shows that nearly all f the posterior probability is at the lower bound of $\eta = 1$. This suggests that the posterior might have selected an even weaker heuristic, if that option had been available.

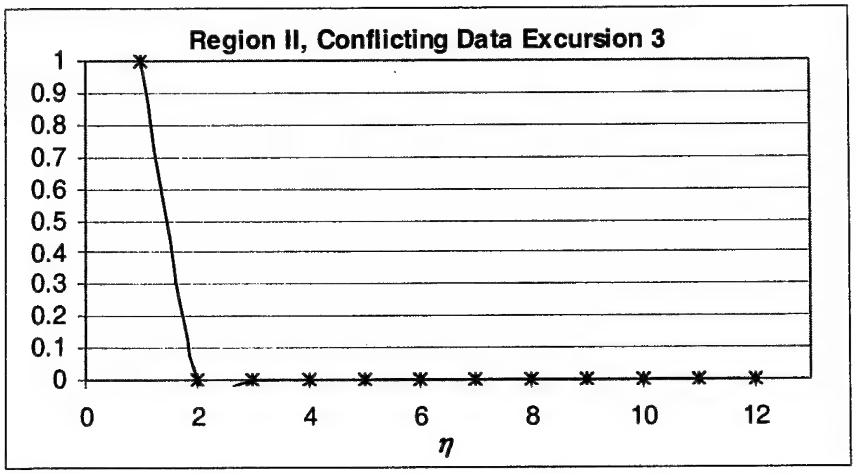

Figure 26. Marginal Posterior Distribution for η
from Excursion 3 in Region II.b

Excursion 4.) Observe the data $w[1] = 4.18325207 = z[59]$, $v_1 = 1$; $w[2] = 4.19713248 = z[60]$, $v_2 = 25$. This sample would be expected to produce a posterior distribution like that of Excursion 3. Figure 27 lives up to this expectation.

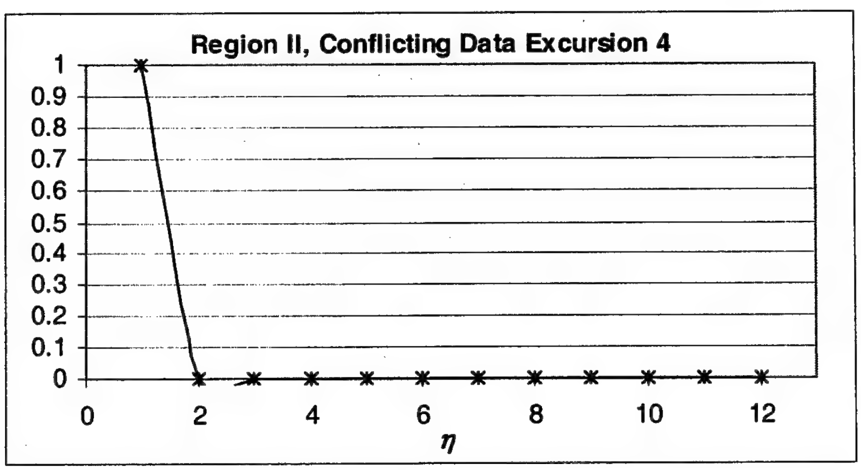

Figure 27. Marginal Posterior Distribution for η
from Excursion 4 in Region II.b

5.4 Region III Excursions

Region III consists of situations where the practitioner has little external knowledge about the problem class but detailed external knowledge about the heuristic class. In these situations a disperse prior is used for the problem class with a tight prior used for the heuristic class. In each of our excursions we will use a relatively disperse problem class prior and a fixed heuristic class prior. As a result, the marginal posterior distributions we will investigate are the ones for (μ, σ) and $Z[1]$.

5.4.1 Problem Class Prior Models

For our Region III excursions we will use a doubly truncated normal model with $\psi = 60$, $LB = 0$, and $UB = 4.71$. We will use the same four prior models for the hyperparameters (μ, σ) discussed in Section 4.7.2.1:

(1.) (μ, σ)~Uniform[$\mu L=LB$, $\mu U=UB$]×Uniform[$\sigma L=0.1$, $\sigma U=15.1$],

(2.) (μ, σ)~Uniform[$\mu L=LB$, $\mu U=UB$]×(K/σ on [$\sigma L=0.1$, $\sigma U=15.1$]),

(3.) (μ, σ)~

$$\left(K\left(\frac{(UB-\mu)}{(UB-LB)}\right)\left(1-\frac{(UB-\mu)}{(UB-LB)}\right) \text{ on } \left[\mu L = LB, \mu U = UB\right]\right)\times\text{Uniform}[\sigma L=0.1,\ \sigma U=15.1],$$

(4.) (μ, σ)~$\left(K\left(\frac{(UB-\mu)}{(UB-LB)}\right)\left(1-\frac{(UB-\mu)}{(UB-LB)}\right) \text{ on } \left[\mu L = LB, \mu U = UB\right]\right)\times(K/\sigma$ on [$\sigma L=0.1$, $\sigma U=15.1$]), where K represents proportionality constants that will not be included explicitly, but become part of the final normalization process.

5.4.2 Heuristic Class Prior Models

We use the same fixed prior models for $\mathcal{H}_{\text{PRS}}$ and $\mathcal{H}_{\min\{12\ \text{PRS}\}}$ in Region III as were used in Region I excursions. These priors are of the form $p_i = \left(\frac{\psi - i + 1}{\psi}\right)^{\eta} - \left(\frac{\psi - i}{\psi}\right)^{\eta}$ with $\eta = 1$ for $\mathcal{H}_{\text{PRS}}$ and $\eta = 12$ for $\mathcal{H}_{\min\{12\ \text{PRS}\}}$.

5.4.3 Supporting Data Excursions for Region III.a

Excursion 1.) Using the fixed heuristic class prior for $\mathcal{H}_{\text{PRS}}$, observe the data $w[1] = 3.29001143$, $v_1 = 1$; $w[2] = 3.58874169$, $v_2 = 1$, which are actual observations from two realizations of $\mathcal{H}_{\text{PRS}}$. As was the case when this data was presented in Region IV.a, we would like for our problem class posterior to favor (μ, σ) near (3.2, 0.6). We would also like for the posterior to favor $Z[1] < w[1]$, since the optimal value is actually at 2.20484126. Although we have a fixed heuristic class model that accurately reflects our sample, we are still using a disperse problem class model, so a sample of size two could very well produce misleading results.

The marginal posterior on (μ, σ) is tighter than the prior distribution under all four prior models, yet it is not as tight as we have seen in most of our previous excursions. Under all four prior models, the grid point (3.46, 0.1) garners a significant amount of the posterior probability, from 18% under Prior Model (1.) to 64% under Prior Model (4.). This grid point is the mode under all except for Prior Model (1.), which has its mode at (4.71, 0.6) with just under 20% of the posterior probability. The rest of the high probability grid points vary notably under each prior, reflecting the structure of the prior model. For instance, Prior Model (2.) and Prior Model (4.) strongly favor grid points with $\sigma = 0.1$ or $\sigma = 0.6$ due to their $1/\sigma$ term. Prior Model (3.) and Prior Model (4.) favor grid points near the middle of [0, 4.71], due to their structure.

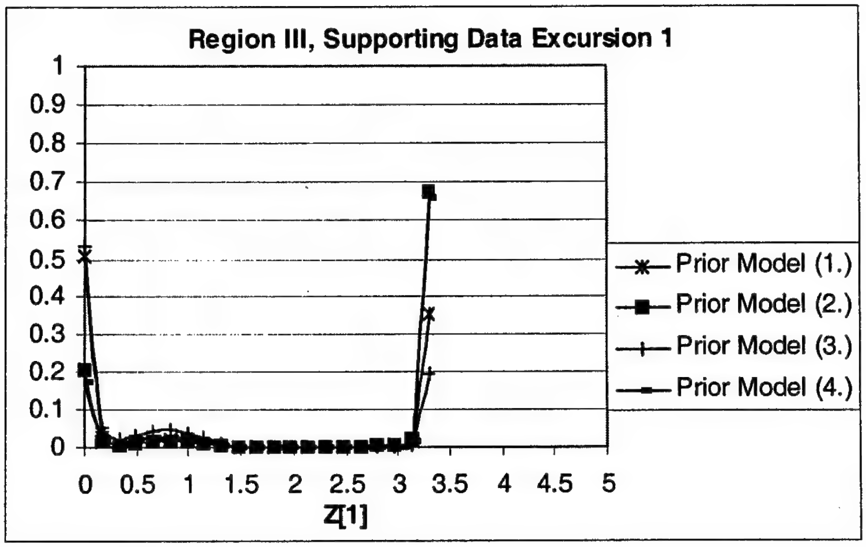

Figure 28. Marginal Posterior Distribution for
$Z[1]$ from Excursion 1 in Region III.a

The marginal posterior distribution for $Z[1]$ plotted in Figure 28 is interesting in several respects. First, we note the distinct modes at both boundaries, $w[1]$ and $LB = 0$. The form of the problem class prior model plays a large role in determining which of these endpoints get most of the posterior probability. $Z[1] = w[1]$ gets more posterior probability under Prior Model (2.) and Prior Model (4.) due to how their $1/\sigma$ term favors a tight distribution of solution values for problem instances from this class. Prior Model (1.) and Prior Model (3.) do not have this tendency toward tightly clustered solution values, so $Z[1] = LB$ gets the majority of the posterior probability under these priors. Also, Prior Model (3.) and Prior Model (4.) show a slightly elevated posterior probability for Z[1] in [0.5, 1.25], apparently corresponding to the term in their prior that favors a problem class mean near the middle of [0, 4.71].

Excursion 2.) Using the fixed heuristic class prior for $\mathcal{H}_{\min\{12\text{ PRS}\}}$, observe the data $w[1] = 2.22373805$, $v_1 = 1$; $w[2] = 2.39441634$, $v_2 = 1$, which are actual observations from two realizations of $\mathcal{H}_{\min\{12\text{ PRS}\}}$. In this excursion, we would again like for the posterior to favor values of (μ, σ) near (3.2, 0.6), based upon our knowledge of the solution value histogram for the enumeration of this problem instance. This sample is considerably stronger than that presented in Excursion 1, but we still might hope that the marginal

posterior distribution for $Z[1]$ favors a value lower than $w[1]$, since the true optimal value is at 2.20484126. However, since our sample is very small and our prior model is disperse, there is a possibility that the posterior will again be misleading in some way.

Under all four prior models the mode of the marginal posterior occurs at $(\mu, \sigma) =$ (3.46, 0.6), with 19-24% of the posterior probability at this grid point. The second-highest probability grid point is (3.62, 0.6) under all four prior models, with the pair of grid points holding a combined 37-50% of the probability in the marginal posterior. The grid points (3.77, 0.6) and (3.30, 0.6) are also among the five highest probability grid points in the marginal posterior distribution under all four prior models.

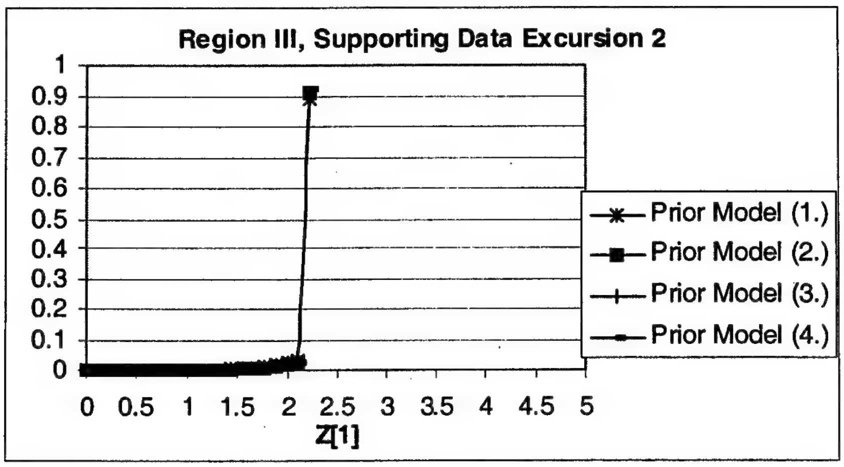

Figure 29. Marginal Posterior Distribution for
Z[1] from Excursion 2 in Region III.a

As we observed in the Region IV excursions, a relatively disperse problem class prior distribution allows the posterior to favor problem class hyperparameter values that are sufficiently high and tight to suggest that we have observed the optimal value in our sample. The results of Excursion 2 in Region III.a show that having a fixed heuristic class model does not eliminate this phenomenon.

5.4.4 Conflicting Data Excursions for Region III.b

Excursion 1.) Using the fixed heuristic class prior for $\mathcal{H}_{PRS}$, observe the data $w[1] =$ 2.22373805, $v_1 = 1$; $w[2] = 2.39441634$, $v_2 = 1$, which are actual observations from two

realizations of $\mathcal{H}_{\min\{12\ PRS\}}$. Since our prior model is for a heuristic that is actually weaker than the sample we are presenting, we would expect the posterior distribution to be misled into favoring a notably higher value for μ and a smaller value for σ than (3.2, 0.6).

Interestingly, the marginal posterior on (μ, σ) has the same three highest-probability grid points as in Excursion 2 in Region III.a, which presented the same data under a different heuristic prior model. In both excursions, the mode of the marginal posterior on (μ, σ) is at (3.46, 0.6), and the next highest probability grid points are at {(3.62, 0.6), (3.30, 0.6)} under Prior Models (1.), (2.), (3.) and at {(3.62, 0,6), (2.52, 0.1)} under Prior Model (4.). However, this excursion features a posterior that is more disperse than before, with the mode capturing only 14-22% of the posterior probability and the two highest probability grid points in the posterior accounting for only 28-40% of the total probability. The analogous percentages from the previous excursion were 19-24% and 37-50%. This suggests that use of an incorrect heuristic prior model can dilute the impact of the sample data on the problem class posterior distribution.

Figure 30 shows that the marginal posterior distribution on $Z[1]$ is also slightly different here than when the same sample was presented to a prior that had $\eta = 12$, for the correct heuristic model. Both this plot and the plot from Region III.a, Excursion 2 have a large spike at $Z[1] = w[1]$. However, this posterior also shows an increased probability at and just above the lower bound of zero and in the range [1.75, 2.00].

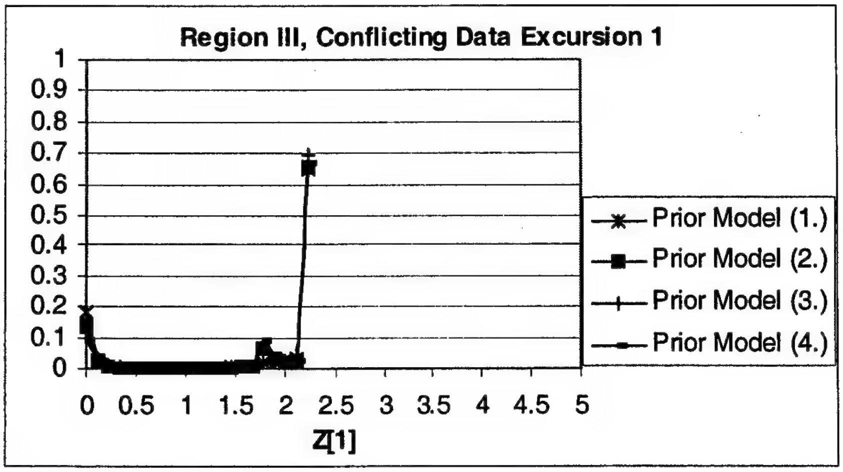

Figure 30. Marginal Posterior Distribution for
$Z[1]$ from Excursion 1 in Region III.b

Excursion 2.) Using the fixed heuristic class prior for $\mathcal{H}_{\min\{12\text{ PRS}\}}$, observe the data $w[1] = 3.29001143$, $v_1 = 1$; $w[2] = 3.58874169$, $v_2 = 1$, which are actual observations from two realizations of $\mathcal{H}_{\text{PRS}}$. Since our prior model is for a heuristic that is actually stronger than the sample we are presenting, we would expect the posterior distribution to be misled into favoring a notably higher value for μ and perhaps a larger value for σ than (3.2, 0.6).

In fact, although the marginal posterior for (μ, σ) strongly favors a much higher value of μ, it again tends to select $\sigma = 0.6$. Under Prior Model (1.) and Prior Model (2.) the mode was at (4.71, 0.6), with this grid point accounting for 67-69% of the posterior probability. Because Prior Model (3.) and Prior Model (4.) have a term that favors μ near the middle of [0, 4.71], the posterior mode under these priors was at (4.56, 0.6), with 68% and 51% of the posterior probability. This grid point is also the second highest probability grid point under Prior Model (1.) and Prior Model (2.), with 25% of the posterior probability under both priors. Under Prior Model (3.), the second highest probability grid point is (4.40, 0.6) with 23% of the posterior probability. Prior Model (4.) has 28% of its posterior probability at its second highest probability grid point, (3.77, 0.1), reflecting the combined effect of the $1/\sigma$ term and the term favoring μ in the middle of [0, 4.71].

As shown in Figure 31, the marginal posterior distribution on $Z[1]$ shows a very strong spike at $Z[1] = w[1]$. This is another indication that using a fixed value of $\eta = 12$ for

the heuristic class prior model has a considerable influence upon the posterior distribution, since the analogous plot for $\eta = 1$ in Figure 28 shows a much larger probability of $Z[1] < w[1]$.

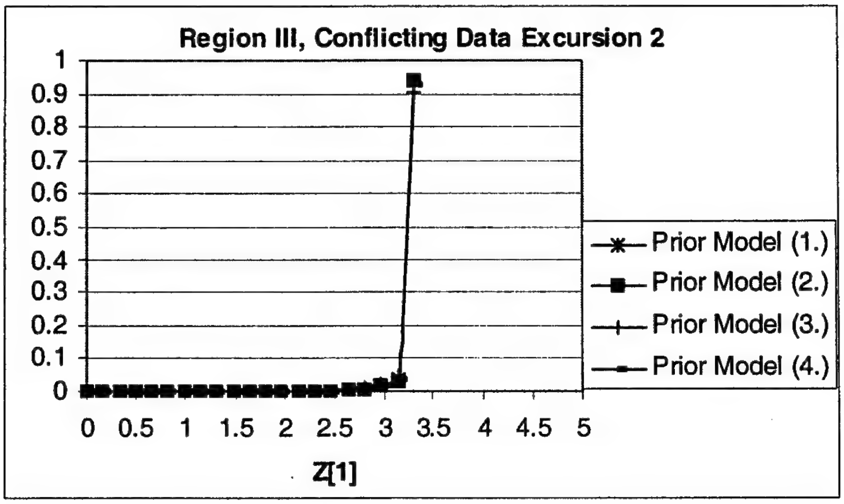

Figure 31. Marginal Posterior Distribution for $Z[1]$ from Excursion 2 in Region III.b

Excursion 3.) Using the fixed heuristic class prior for $\mathcal{H}_{\text{PRS}}$, observe the data $w[1] = 2.20872027$, $v_1 = 1$; $w[2] = 2.22373805$, $v_2 = 1$, which are actual observations from two realizations of $\mathcal{H}_{\min\{25\text{ PRS}\}}$. In this excursion we expect to see behavior like that in Excursion 1 of Region III.b, since both of these excursions feature a fixed heuristic prior model with $\eta = 1$ and a sample that actually results from a much stronger heuristic. However, since this sample is from an even stronger heuristic than that in Excursion 1 of Region III.b, we might expect the posterior to favor μ even lower than before and σ even smaller than before.

Indeed, this is exactly what occurs in the marginal posterior on (μ, σ). The posterior mode is at (2.36, 0.1) under all four prior models, with this grid point capturing 32-55% of the posterior probability. The second highest probability grid point is (2.04, 0.1), with 49-83% of the posterior probability going to these two highest probability grid points. Figure 32 reveals a marked similarity to Figure 30. Both plots have a substantial

spike at $Z[1] = w[1]$ along with elevated probability at $Z[1] = LB = 0$ and $Z[1]$ on [1.75, 2.0]. This plot has more probability on [1.75, 2.0] and less at $Z[1] = 0$ than the earlier plot.

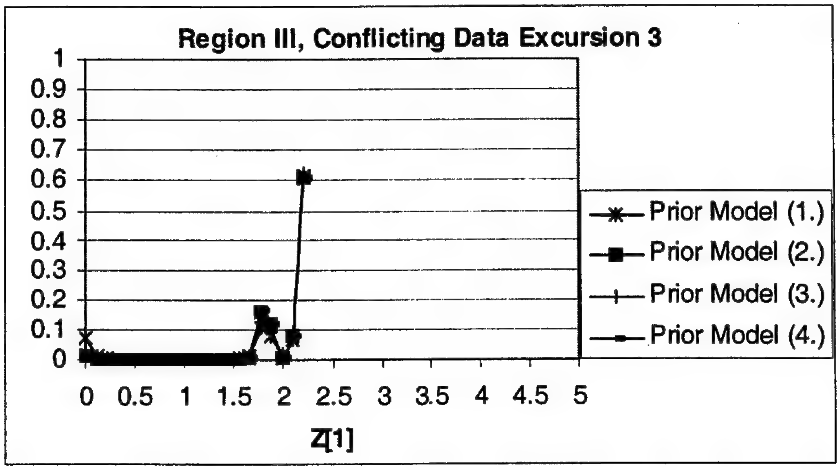

Figure 32. Marginal Posterior Distribution for
Z[1] from Excursion 3 in Region III.b

Excursion 4.) Using the fixed heuristic class prior for $\mathcal{H}_{\min\{12\text{ PRS}\}}$, observe the data $w[1] = 2.20872027$, $v_1 = 1$; $w[2] = 2.22373805$, $v_2 = 1$, which are actual observations from two realizations of $\mathcal{H}_{\min\{25\text{ PRS}\}}$. As with Excursion 1, we would expect the posterior distribution to be misled into favoring a higher value for μ and a smaller value for σ than (3.2, 0.6) since our prior model is for a heuristic that is actually weaker than the sample we are presenting.

The marginal posterior distribution on (μ, σ) has its mode at (2.36, 0.1) under all four prior models. This grid point captured 43-78% of the posterior probability. Under Prior Model (1.) and Prior Model (3.), the second highest probability grid point is at (3.30, 0.6) with about 9% of the posterior probability at this grid point. Due to the influence of $1/\sigma$ term, Prior Model (2.) and Prior Model (4.), have the second highest probability grid point is at (2.52, 0.1) with about 13% of the posterior probability at this grid point. Instead of shifting towards a higher value for μ, the posterior distribution favors a very tight problem instance.

Figure 33 shows the marginal posterior distribution on Z[1]. Not surprisingly, we see a huge spike at $Z[1] = w[1]$.

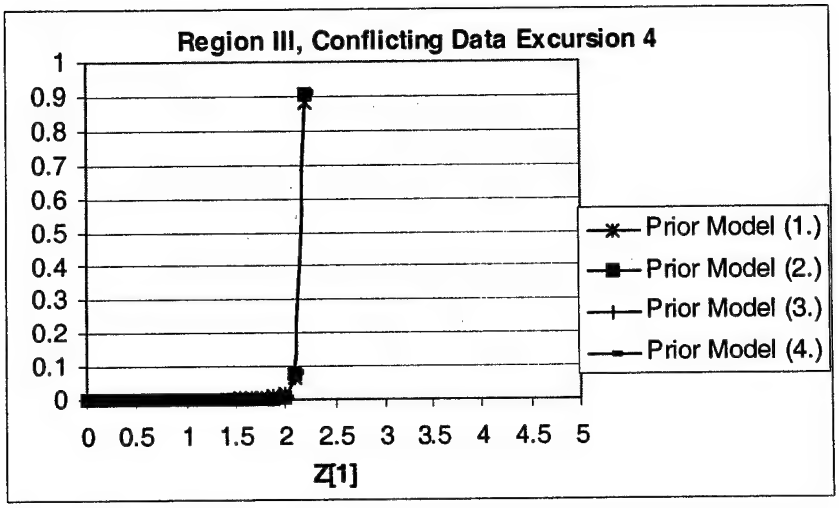

Figure 33. Marginal Posterior Distribution for
$Z[1]$ from Excursion 4 in Region III.b

Excursion 5.) Using the fixed heuristic class prior for $\mathcal{H}_{PRS}$, observe the data $w[1] = 4.18325207 = z[59]$, $v_1 = 1$; $w[2] = 4.19713248 = z[60]$, $v_2 = 25$. We would like to see our posterior distribution give some indication that our sample violates the assumption that p_i strictly increases with decreasing i, but it is not clear how that would occur. Since this excursion assumes $\eta = 1$, the problem class model could have (μ, σ) about anywhere in their range. However, larger values of μ and smaller values of σ seem more likely.

In fact, the marginal posterior distribution on (μ, σ) has its mode at (4.40, 0.1) under all four prior models, with this grid point capturing 48-96% of the posterior probability. The second highest probability grid point is at (3.93, 0.1) with the two points combining for 50-98% of the posterior probability.

The marginal posterior distribution on $Z[1]$is shown in Figure 34. This plot shows a pattern that we have seen in many of the excursions that assumed a fixed heuristic prior model with $\eta = 1$. It has both a spike at $Z[1] = w[1]$ and a spike at $Z[1] = LB = 0$. As in previous cases with $\eta = 1$, due to their tendency toward small values of σ, Prior Model (2.) and Prior Model (4.) have more posterior probability at $Z[1] = w[1]$ than do the other two models.

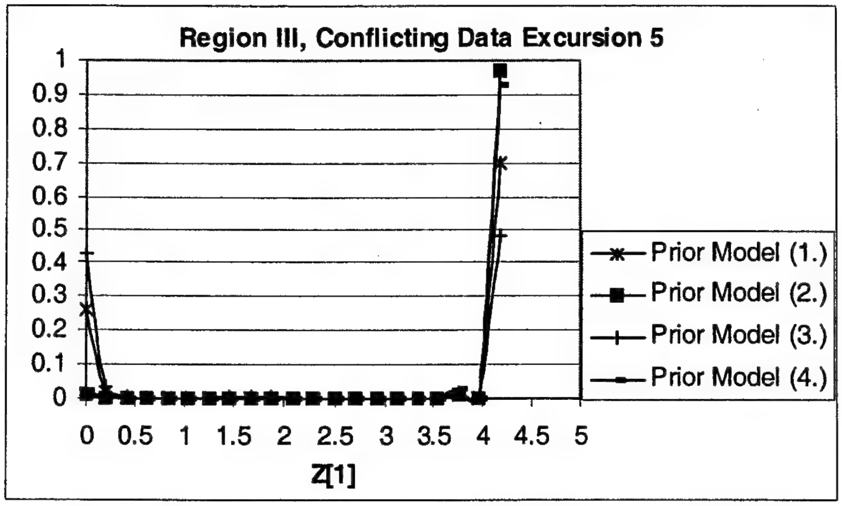

Figure 34. Marginal Posterior Distribution for $Z[1]$ from Excursion 5 in Region III.b

Excursion 6.) Using the fixed heuristic class prior for $\mathcal{H}_{\min\{12\text{ PRS}\}}$, observe the data $w[1] = 4.18325207 = z[59]$, $v_1 = 1$; $w[2] = 4.19713248 = z[60]$, $v_2 = 25$. We would expect this excursion to give results very much like those in Excursion 4 of Region IV.b, since that excursion featured the same data and its posterior favored $\eta = 12$. In Excursion 4 of Region IV.b, the posterior strongly favors a problem class with $(\mu, \sigma) = (4.40, 0.1)$, and the marginal posterior on $Z[1]$ has a very strong spike at $Z[1] = w[1]$. Indeed, this excursion also has a marginal posterior distribution for (μ, σ) with over 99.9% of the posterior probability at (4.40, 0.1), and Figure 35 shows a marginal posterior distribution on $Z[1]$ with almost all of its probability in the spike at $Z[1] = w[1]$.

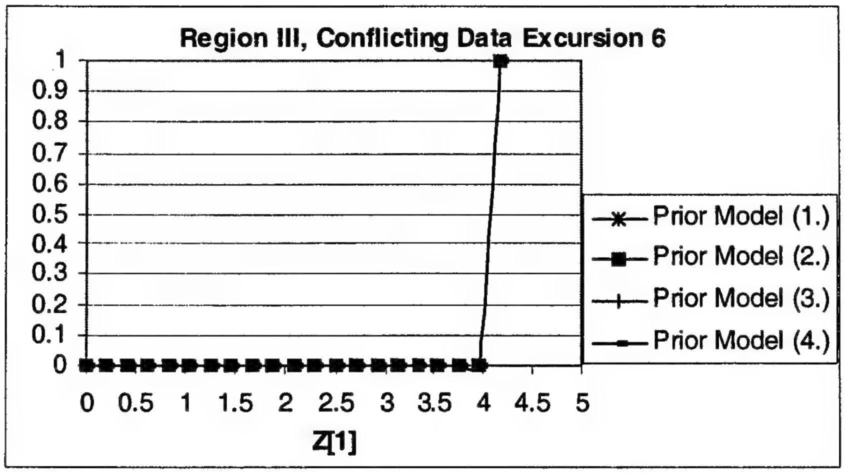

Figure 35. Marginal Posterior Distribution for
$Z[1]$ from Excursion 6 in Region III.b

5.5 Additional Scenario Excursions

In many of the previous excursions we see how dispersion in the problem class prior model can allow for a misleading posterior distribution, particularly when the sample is very small. Although the feasible parameter space in the Region IV excursions included $\eta = 1$, the posterior distributions rarely allocated much probability to this value, even when the data presented was actually the result of Pure Random Sampling. Our hypothesized explanation for this behavior is that the PRS sample presented contained only two observations and these observations were relatively close together with respect to some combinations of the problem class hyperparameters. This led the posterior to favor a stronger heuristic by selecting an appropriate combination of problem class hyperparameters. In order to explore this hypothesis, we investigated another set of excursions. These excursions fall near the center of Figure 4 and Figure 5, because both problem class and heuristic class prior models are somewhat disperse.

5.5.1 Problem Class Prior Model

In this set of excursions we fixed all of the hyperparameters of the truncated normal problem class model. We selected our problem class prior model by using Excel to compare plots of truncated normal densities with various hyperparameter settings to the histogram that resulted from enumeration of our 6-node Euclidean TSP instance with

minisum objective. As before, we used $\psi = 60$, $LB = 0$, and $UB = 4.71$. Based on the Excel comparisons, we selected $\mu = 3.2$ and $\sigma = 0.6$ to complete our problem class prior model.

5.5.2 Heuristic Class Prior Model

As in our Region II and Region IV excursions, we used a relatively disperse heuristic class prior. This prior was formed by applying $p_i = \left(\frac{\psi - i + 1}{\psi}\right)^\eta - \left(\frac{\psi - i}{\psi}\right)^\eta$ with $\eta = \{1, 2, \ldots, 12\}$.

5.5.3 Supporting Data Excursions

Excursion 1.) Observe the data $w[1] = 3.29001143$, $v_1 = 1$; $w[2] = 3.58874169$, $v_2 = 1$, which are actual observations from two realizations of $\mathcal{H}_{PRS}$. We hope that using fixed values for (μ, σ) that fairly accurately reflect this problem instance will result in a posterior distribution that favors $\eta = 1$ and $Z[1] < w[1]$ for this sample. Figure 36 and Figure 37 show the results.

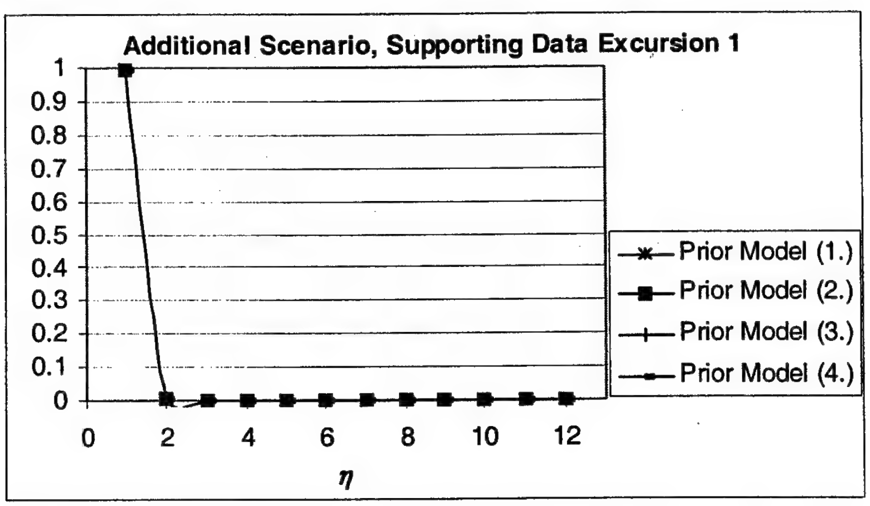

Figure 36. Marginal Posterior Distribution for η from Excursion 1 in Additional Scenario a

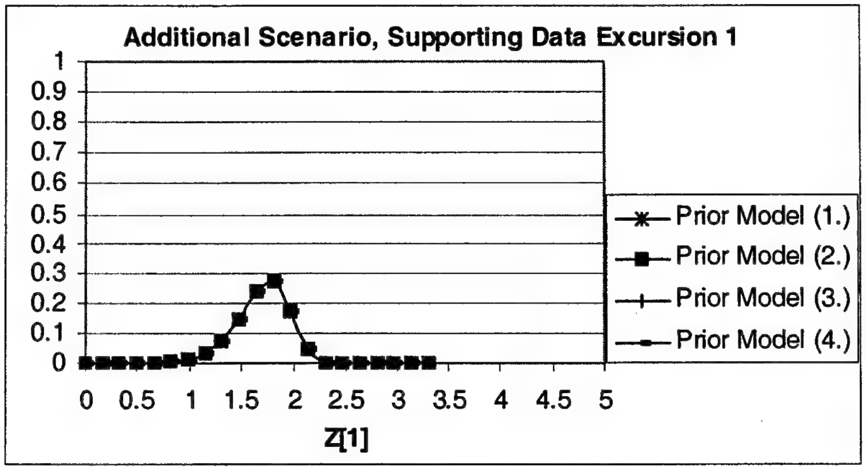

Figure 37. Marginal Posterior Distribution for
Z[1] from Excursion 1 in Additional Scenario a

The marginal posterior on η in Figure 36 is as we hoped. Also, Figure 37 reveals that the marginal posterior on Z[1] does indeed have most of its probability substantially below Z[1]. In fact, there is virtually no probability at Z[1] = w[1]. Instead, the plot looks much like we would expect the density of the minimum of 60 random variables from the N(0, 4.71, 3.2, 0.6) problem class model. The lack of a noticeable spike at Z[1] = w[1] is the result of w[1] being far into the upper tail of this distribution.

Excursion 2.) Observe the data w[1] = 2.22373805, v_1 = 1; w[2] = 2.39441634, v_2 = 1, which are actual observations from two realizations of $\mathcal{H}_{\min\{12\ \mathrm{PRS}\}}$. Again, we expect the fixed values of (3.2, 0.6) for the problem class hyperparameters to lead to a posterior distribution that more accurately favors the true parameter values, that is, η = 12 and Z[1] around 2.20. Figure 38 and Figure 39 show that the results essentially match with our expectations.

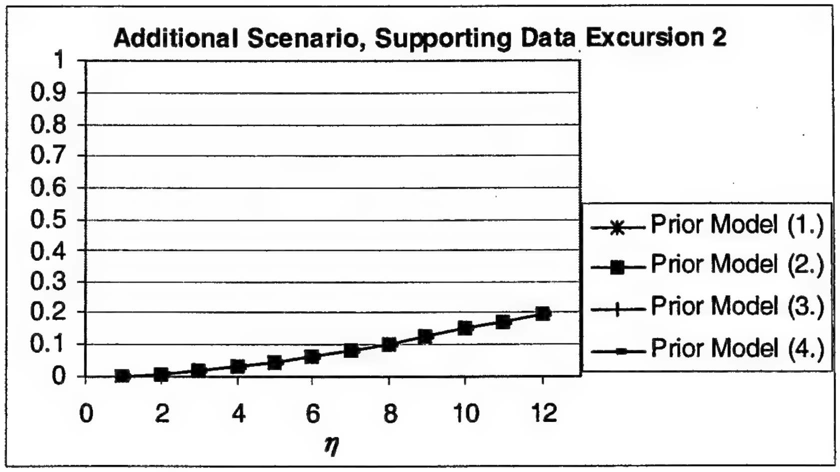

Figure 38. Marginal Posterior Distribution for η from Excursion 2 in Additional Scenario a

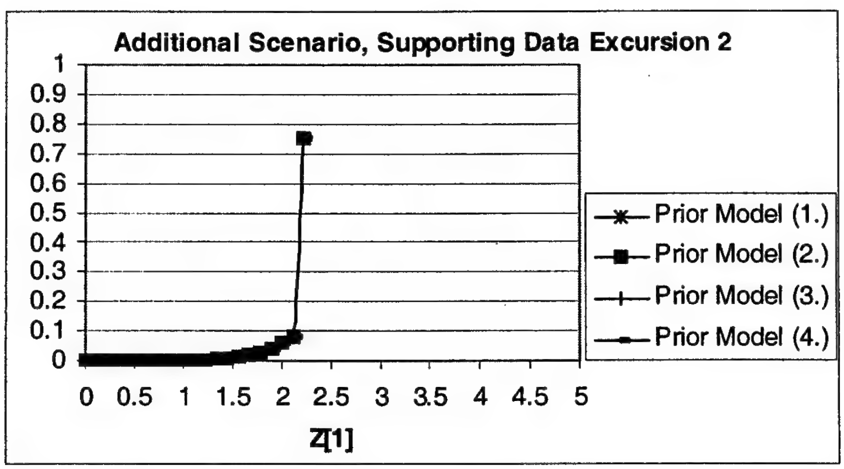

Figure 39. Marginal Posterior Distribution for $Z[1]$ from Excursion 2 in Additional Scenario a

Excursion 3.) Observe the data $w[1] = 2.20872027$, $v_1 = 1$; $w[2] = 3.29001143$, $v_2 = 1$; $w[3] = 3.58874169$, $v_3 = 1$; $w[4] = 4.18325207$, $v_4 = 1$, which were carefully selected to "look" like a PRS sample from the specified problem class model. As in Excursion (1.) and Excursion (2.) above, we expect the results to better reflect what we see as the true nature of the heuristic. Since this sample is fairly spread out relative to the problem class

hyperparameters, this should translate into a posterior distribution that favors $\eta = 1$ and $Z[1]$ near 2.0. Figure 40 and Figure 41 show plots that match with these expectations.

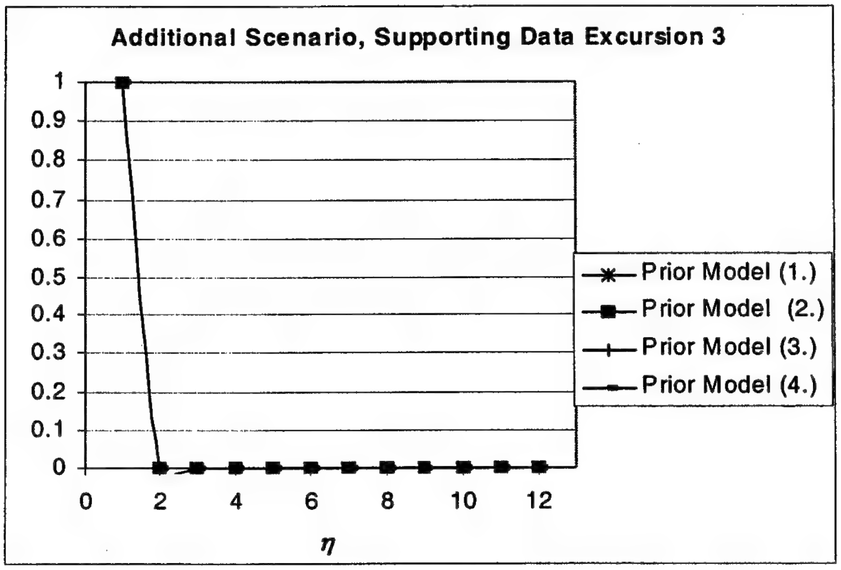

Figure 40. Marginal Posterior Distribution for η from Excursion 3 in Additional Scenario a

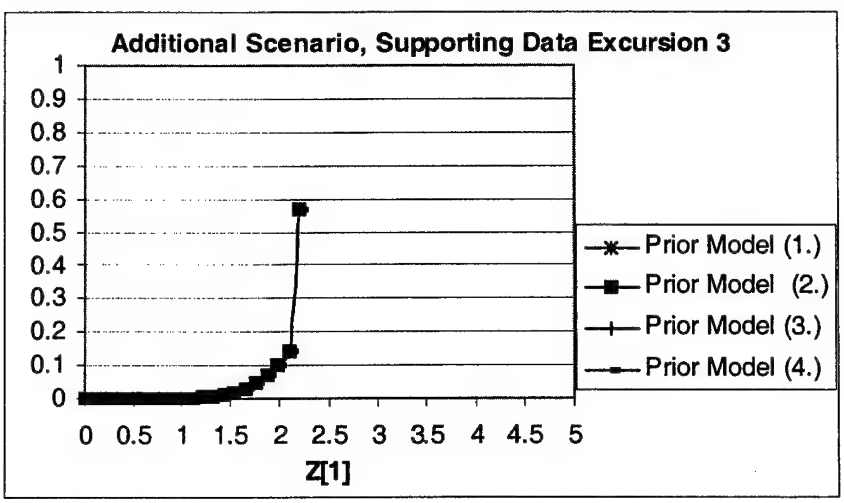

Figure 41. Marginal Posterior Distribution for $Z[1]$ from Excursion 3 in Additional Scenario a

5.5.4 Conflicting Data Excursion

Excursion 1.) Observe $w[1] = 4.18325207 = z[59]$, $v_1 = 1$; $w[2] = 4.19713248 = z[60]$, $v_2 = 1$. Since the solution values in this sample are far into the upper tail of the problem class model, we would expect for the posterior to favor $\eta = 1$, which is the weakest heuristic model option available in the prior. As in supporting Excursion 1 for this scenario, we should also see a marginal posterior on $Z[1]$ that reflects the distribution of the minimum of 60 observations from N(0, 4.71, 3.2, 0.6). Figure 42 and Figure 43 show precisely this behavior.

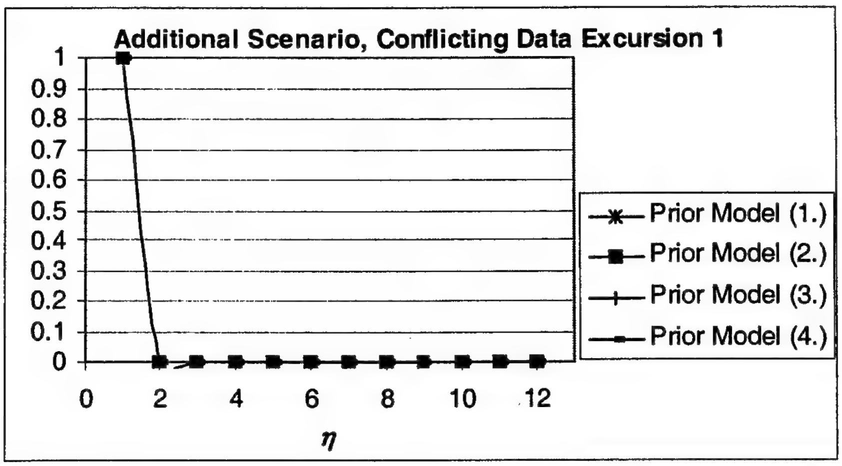

Figure 42. Marginal Posterior Distribution for η from Excursion 1 in Additional Scenario b

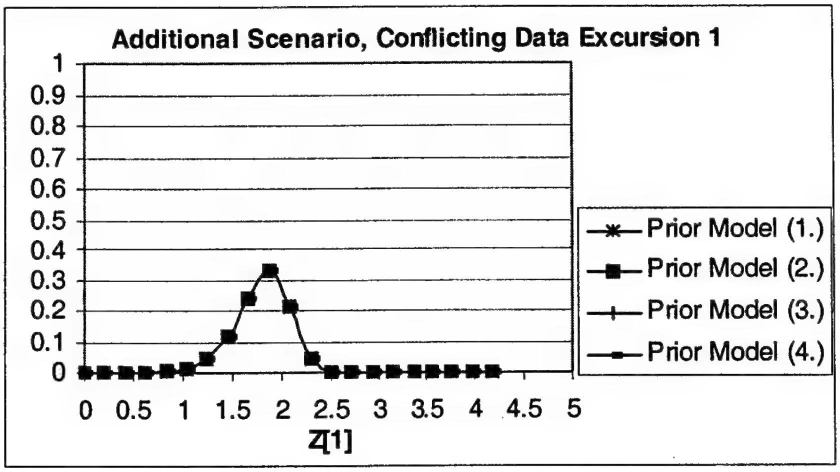

Figure 43. Marginal Posterior Distribution for
Z[1] from Excursion 1 in Additional Scenario b

5.5.5 Summary of Additional Scenario Results

The results of excursions on the additional scenario show that even a somewhat disperse problem class model (i.e. one with fixed hyperparamters) is much more resistant to the tendency to indicate an overly strong heuristic. This suggests that future applications of a P-H methodology to heuristic-quality assessment will benefit greatly when the practitioner uses a variety of methods to learn about the problem instance or instances at hand, then incorporates this information into more reasonable and less disperse problem class prior models.

6 SUMMARY AND CONCLUSIONS

6.1 Summary of Empirical Results

Although most of the empirical results described in Chapter 5 matched the expectations discussed in conjunction with Figure 5, there were a few unexpected results that provide important insight. In particular, results of Region IV excursions revealed that the posterior can be misled to favor a strong heuristic even when the data presented are from a weak heuristic or actually in the far upper tail of the problem instance solution values. This occurrence shows the importance of the prior distributions. Although the discrete uniform distribution on $\{1, 2, \ldots, 12\}$ used for η was intended to be very disperse, it implicitly favors a stronger heuristic in that most of these twelve equally likely values for η reflect a strong heuristic. When this heuristic prior model was used in conjunction with the disperse problem class priors of Region IV, the posterior was able to select a combination of problem class hyperparameters, (μ, σ), that would allow weak data to appear as though they resulted from a strong heuristic.

While the empirical results illustrate the potential for a meaningful Bayesian inference approach using the P-H Probability Model, they also underscore the importance of carefully formulating the prior models and considering their postential effects. In particular, the practitioner should realize that prior models intended to be very disperse may contain implicit biases. On the other hand, overly disperse prior models (particularly for the problem class) may be too easily misled by small or "unlucky" samples.

6.2 Benefits of a P-H Methodology for Heuristic-Quality Assessment

The primary benefit of using a P-H methodology to assess heuristic quality is that the P-H Probability Model accurately reflects the structure involved when heuristic techniques are applied to solution of combinatorial optimization problems. As a result, it is more generally applicable to the heuristic assessment process than most previous methods.

For instance, it can handle samples showing discreteness, without incorporating special workarounds like that proposed by Los and Lardinois (1982). Moreover, the P-H methodology can be easily applied to address the perspective of the heuristic user with multiple heuristics. Unlike in Dannenbring's (1977) efforts to incorporate multiple heuristics, an approach based upon the P-H Probability Model would not be limited to point estimation. Finally, the P-H Probability Model can be extended across multiple instances from a problem class to allow the heuristic researcher to assess heuristic performance across the class.

In order to employ the P-H Probability Model in assessing heuristic quality, the heuristic practitioner must think carefully about the problem instance, or the problem class, and the heuristic class. While this process of model building for the problem class and heuristic class priors may be difficult and intimidating, working through the process leads the practitioner to develop a better understanding of the structure underlying the observed solution values. Moreover, being forced to clearly express how much or how little he knows about problem and heuristic structure gives the practitioner a more accurate appreciation for the assessment results.

In contrast, most previous approaches to statistical assessment of heuristic quality offer a method that proposes to work well in the limit for virtually any problem-heuristic context. These one-size-fits-all approaches can often lull the practitioner into a false sense of security, providing him with little insight into the process and little means for critical evaluation of the results.

6.3 Drawbacks of a P-H Methodology for Heuristic-Quality Assessment

Unfortunately, application of a P-H methodology is not yet practical in general. The number of terms in the sums for the marginal posterior on $Z[1]$ grows as $\binom{\psi-1}{k-1}+\binom{\psi}{k-1}$, where k is the number of distinct solution values in the sample. For practical applications of the methodology, both ψ and k are likely to be quite large. Before the methodology can be feasibly applied to these realistic situations, better computational methods or approximations must be developed. For example, by considering the nature of

the prior distributions used for problem and heuristic classes, it might be possible to begin the sum with the largest terms and continue through terms of decreasing probability until a specified tolerance has been reached.

Moreover, many practitioners may be unfamiliar with Bayesian inference and uncomfortable applying it. Until more research is done to improve the accessibility and illustrate the efficacy of this approach, it will be difficult to gain acceptance within the heuristic optimization community.

6.4 Areas for Future Research

Many of the areas for future research on a P-H methodology for heuristic-quality assessment have already been suggested in the discussion of benefits and drawbacks in Sections 6.2 and 6.3. The most pressing need in continuing this research is dealing with the computational difficulties that would accompany large values of ψ and k. Until this issue is addressed, more extensive empirical investigation of the sort that would build credibility with heuristic practioners is not practical.

Another research thread that could lead to widespread acceptance of a P-H methodology for heuristic-quality assessment is investigating the nature of prototypical problem classes with an eye towards problem class prior model specification. Patel and Smith (1983) suggest that a Weibull model form may be an appropriate asymptotic model for continuous-variable problems, and central limiting effects suggest the use of a normal model when the objective function is a sum of numerous real numbers. However, identifying these distributional forms is not sufficient guidance. Future research could help practitioners to specify values for the hyperparameters of these models in a straightforward and meaningful fashion.

Once a more efficient computational procedure and more extensive guidance on prior model formulation are developed, one could conduct a series of extensive empirical evaluations and comparisons to previous approaches. This is precisely the type of evidence heuristic practitioners are likely to require before they are convinced of the value of a Bayesian inference approach to heuristic-quality assessment based upon the P-H Probability Model.

LIST OF REFERENCES

LIST OF REFERENCES

C.G.E. BOENDER, A.H.G. RINNOOY KAN (1983). A Bayesian Analysis of the Number of Cells of a Multinomial Distribution, *Statistician* **32**, 240-248.

C.G.E. BOENDER, A.H.G. RINNOOY KAN, L. STOUGIE, G.T. TIMMER (1982). A Stochastic Method for Global Optimization, *Mathematical Programming* **22**, 125-140.

D. J. CLOUGH (1969). An Asymptotic Extreme-Value Sampling Theory for Estimation of a Global Maximum, *Canadian Operational Research Society* **7**, 102-115.

S.G. COLES, E.A. POWELL (1996). Bayesian Methods in Extreme Value Modelling: A Review and New Developments, *International Statistical Review* **64**, 119-136.

D. DANNENBRING (1977). Estimating Optimal Solutions for Large Combinatorial Problems, *Management Science* **23**, 1273-1283.

U. DERIGS (1985). Using Confidence Limits for the Global Optimal in Combinatorial Optimization, *Operations Research* **33**, 1024-1049.

R. FISHER, L. TIPPETT (1928). Limiting Forms of the Frequency Distribution of the Largest or Smallest Member of a Sample, *Proceedings of the Cambridge Philosophical Society* **24**, 180-190.

B.L. GOLDEN, F.B. ALT (1979). Interval Estimation of a Global Optimum for Large Combinatorial Problems, *Naval Research Logistics Quarterly* **26**, 69-77.

E.J. GUMBEL (1958). *Statistics of Extremes*, Columbia University Press, New York.

M. LOS, C. LARDINOIS (1982). Combinatorial Programming, Statistical Optimization and the Optimal Transportation Network Problem, *Transportation Research* **16B**, 89-124.

E.L. LAWLER, J.K. LENSTRA, A.H.G. RINNOOY KAN, D.B. SHMOYS (1985). *The Traveling Salesman Problem*: A Guided Tour of Combinatorial Optimization. John Wiley and Sons, Chichester.

K.L. McROBERTS (1971). A Search Model for Evaluating Combinatorially Explosive Problems, *Operations Research* **19**, 1331-1349.

R.L. NYDICK, JR., H.J. WEISS (1994). An Analytical Evaluation of Optimal Solution Value Estimation Procedures, *Naval Research Logistics* **41**, 189-202.

I.M. OVACIK, S. RAJAGOPALAN, R. UZSOY (2000). Integrating Interval Estimates of Global Optima and Local Search Methods for Combinatorial Optimization Problems, *Journal of Heuristics* **6**, 481-500.

N.R. PATEL, R.L. SMITH (1983). The Asymptotic Extreme Value Distribution of the Sample Minimum of a Concave Function under Linear Constraints, *Operations Research* **31**, 789-794.

M.G. PILCHER, R.L. RARDIN (1987). A Random Cut Generator for Symmetric Traveling Salesman Problems with Known Optimal Solutions, *University Research Initiative in Computational Combinatorics*, Institute for Interdisciplinary Engineering Studies, Purdue University, West Lafayette.

R.L. RARDIN, R. UZSOY (2001). Experimental Evaluation of Heuristic Optimization Algorithms: A Tutorial, *Journal of Heuristics* **7**, 261-304.

S. REITER, G. SHERMAN (1965). Discrete Optimizing, *Journal of the Society for Industrial and Applied Mathematics* **13**, 864-889.

D.S. ROBSON, J. H. WHITLOCK (1964). Estimation of a Truncation Point, *Biometrika* **51**, 33-39.

P. VAN DER WATT (1980). A Note on Estimation of Bounds of Random Variables, *Biometrika* **67**, 712-714.

A.D. WILSON, R.E. KING, J.R. WILSON (2002). The Use of Statistical Estimates of Global Optima as Lower Bounds for the Problem of Scheduling Nonsimilar Groups of Jobs on a Flow Line, Unpublished Report, Department of Industrial Engineering, North Carolina State University, Raleigh, NC.

S. H. ZANAKIS (1979). A Simulation Study of Some Simple Estimators of the Three-Parameter Weibull Distribution, *Journal of Statistical Computing and Simulation* **9**, 419-428.

APPENDICES

Appendix A. A Graphical Comparison of The Truncation-Point Estimators and Zanakis Extreme-Value-Theory Estimator

The graphical representation we develop is a random plot with scaling determined by the distance between the first and second order statistics of the sample, $w[2,n] - w[1,n]$. The horizontal axis represents the location of the single observation, s from the stronger heuristic, relative to $w[1,n]$ and $w[2,n]$. We can view this axis as a representation of the relationships within our solution value data. The vertical axis gives the value of the estimator $\hat{\theta}$ with scale again determined by $w[2,n] - w[1,n]$. This axis may be viewed as depicting the relationship between our estimator and the sample on W.

Note that there is no origin on the plot and the optimal solution value θ is not shown *explicitly*. It is not possible to show the precise value of θ on the plot, since it is a random plot and θ is a fixed value for a given problem instance, located some random distance below $w_{(1)}$. However, as the graphical representation is developed we will indicate areas on the graph where θ cannot be and areas that are feasible for θ.

We first consider the situation where we have collected the sample on W but do not yet know the value of s. At this time, we can form the estimators $\hat{\theta}_{SAMP}$ and $\hat{\theta}_Z$, which is a simple estimator for the Weibull location parameter proposed by Zanakis (1979) and often applied in extreme value theory approaches to optimal value estimation. Note that we may write $\hat{\theta}_Z = \dfrac{w[1,n]w[n,n] - \left(w[2,n]\right)^2}{w[1,n] + w[n,n] - 2w[2,n]}$ as

$$\hat{\theta}_{Z,C} = w[1,n] - \frac{w[2,n] - w[1,n]}{C-2}, \text{ where } C = \frac{w[n,n] - w[1,n]}{w[2,n] - w[1,n]}.$$

Note that C is fixed once the sample on W is known. Moreover, C measures the distance between the $w[n,n]$, the largest member of W, and $w[1,n]$, the smallest member of W, in units of $w[2,n] - w[1,n]$. The value of C indicates the explanatory power of the W sample in estimating the lower bound θ, with large values of C indicating an informative

sample. Note that n has a very direct role in determining C, in that large n will generally produce large C. However, the means of "sampling" W also affects the power of the sample in that a more powerful heuristic will generally result in a probability distribution for W that is more skewed to the left, producing a large value of C.

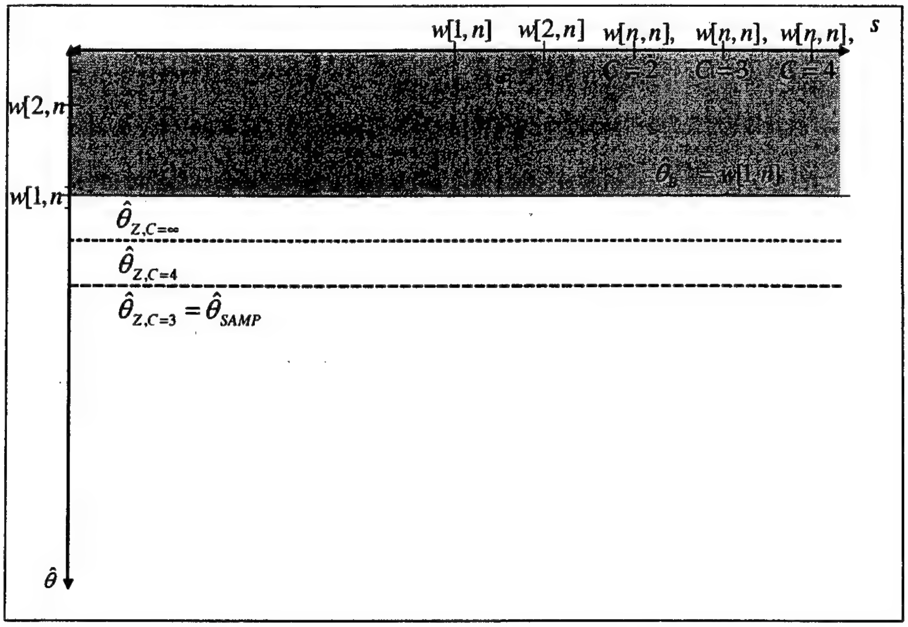

Figure 44. Estimator Diagram Before s is Observed

Figure 44 shows the axes as described along with the available estimators and the boundary on θ provided by $w[1,n]$. The estimate θ cannot lie in the shaded region, since we know that $\theta < w[1,n]$. Hence an estimator $\hat{\theta}$ that enters the shaded region is guaranteed to have positive bias of at least $\hat{\theta} - \hat{\theta}_B$, causing us to treat the heuristic we are evaluating as if it is at least $\hat{\theta} - \hat{\theta}_B$ units closer to the optimal solution value than it really is.

The diagram in Figure 44 shows three possible values of $\hat{\theta}_Z$ rather than three choices for $\hat{\theta}_Z$, since our sample on W determines $\hat{\theta}_Z$ completely. For example, if $w[2,n] = w[1,n]$, we have $C = \infty$ (an extremely informative sample) and $\hat{\theta}_Z$ is the boundary. Not shown on the diagram is the particularly unappealing behavior of $\hat{\theta}_Z$ for $C \leq 2$ (extremely

uninformative samples). As $C \to 2$ from above, $\hat{\theta}_Z \to -\infty$. If $w[n,n] = w[2,n]$, $C = 1$ and $\hat{\theta}_Z = w[2,n] = w[n,n]$, entering the shaded region. Finally, if $w[n,n] = w[2,n] = w[1,n]$, $\hat{\theta}_Z$ has the indeterminate form 0/0.

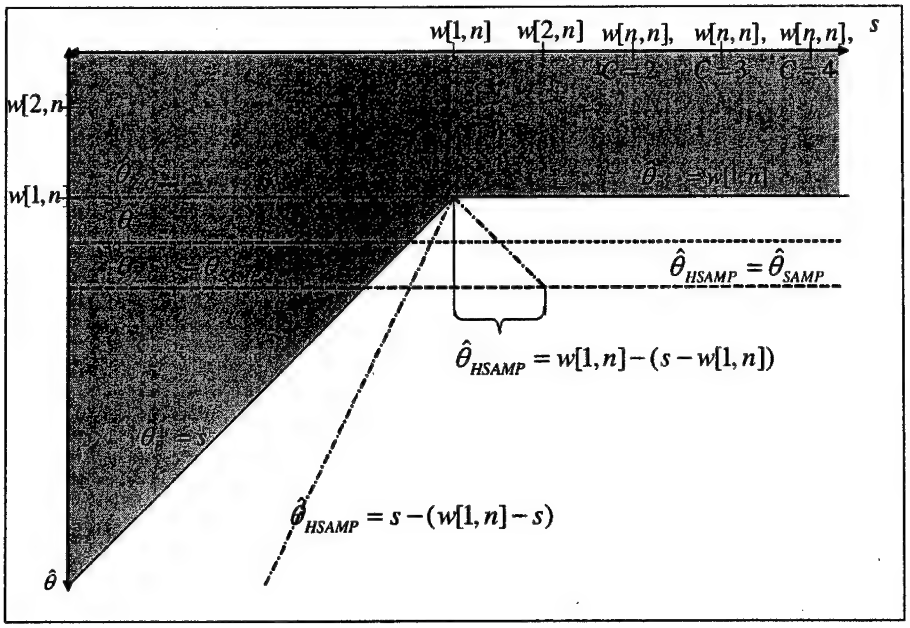

Figure 45. Initial Estimator Diagram Once s is Observed

In Figure 45 we have now made our observation of s. At this point we have a new boundary on θ, $\hat{\theta}_B = \min\{w[1,n], s\}$. The shaded region depicts the resulting new region of infeasibility for θ. We can now write

$$\hat{\theta}_{HSAMP} = \begin{cases} \hat{\theta}_{SAMP}, & \text{if } s > w[2,n] \\ \hat{\theta}_B - (\max\{s, w[1,n]\} - \hat{\theta}_B), & \text{otherwise} \end{cases}.$$

We add $\hat{\theta}_B$ and the $\hat{\theta}_{HSAMP}$ estimator to the plot from Figure 44. Again, we would not want to form estimators of θ that lie in the shaded area.

The plot reveals that where $s < w[1,n]$, $\hat{\theta}_{SAMP}$ and our well-behaved versions of $\hat{\theta}_Z$ have positive bias of at least $\hat{\theta}_Z - s$ where they enter the shaded region. Without incorporating the information in s, these estimators cannot hope to overcome this.

Also note the odd behavior of $\hat{\theta}_{HSAMP}$ for $w[1,n] < s < w[2,n]$. This oddity results from an implicit assumption underlying the mathematics used for incorporating s. In $\hat{\theta}_{HSAMP}$, s is treated as if it were just another observation on W, supplanting $w[1,n]$ or $w[2,n]$ in the estimator if it would replace either of them as an order statistic in the combined sample. However, we know that the combined sample is not an i.i.d. sample, since s is from a stronger heuristic than the one associated with W. The Los and Lardinois (1982) suggestion to use distinct heuristics to provide the n samples of size m for the Weibull estimator would suffer the same weakness—the resulting sample of solution values could not truly be considered samples from the same distribution.

Even if we are not concerned with the theoretical problems associated with $\hat{\theta}_{HSAMP}$, the behavior exhibited in Figure 45 is troubling. Why would we want our estimator to deviate from the boundary on θ in such a non-uniform fashion? It is difficult to conceive of a model for the underlying system (i.e. the solution value space coupled with the behavior of the heuristic) that would justify the selection of such an estimator.

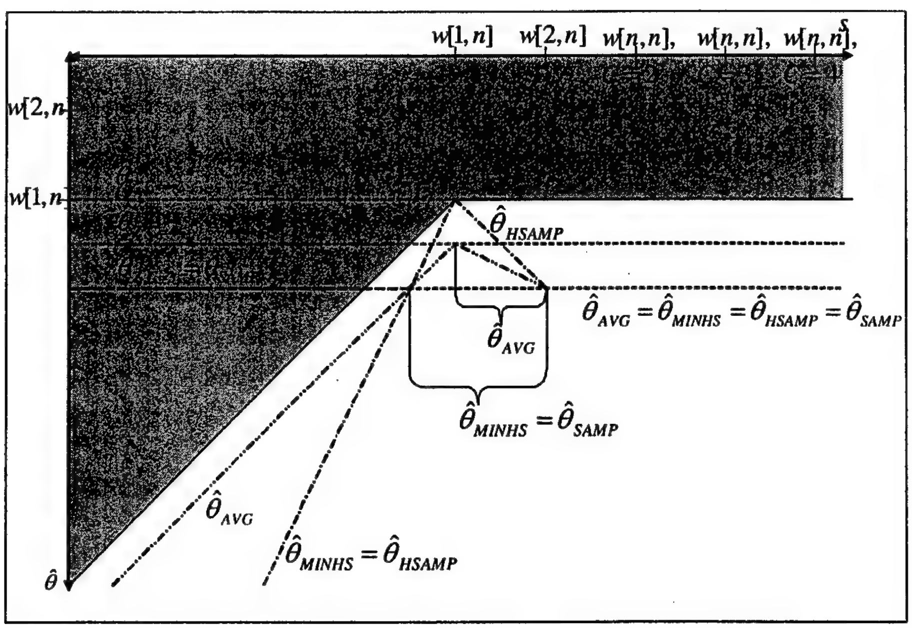

Figure 46. Final Estimator Diagram Once s is Observed

In Figure 46 we have the same information available as in Figure 45. We now form the estimators $\hat{\theta}_{MINHS}$ and $\hat{\theta}_{AVG} = \min\left(w[1,n], s + \frac{w[2,n] - w[1,n]}{2} \right) - (w[2,n] - w[1,n])$. Note that $\hat{\theta}_{MINHS}$ follows $\hat{\theta}_{SAMP}$ for $w[1,n] < s$ and $\hat{\theta}_{HSAMP}$ for $s < w[1,n]$, avoiding the strange behavior of $\hat{\theta}_{HSAMP}$ for $w[1,n] < s < w[2,n]$.

Also note how $\hat{\theta}_{AVG}$ drops at the rate s as does $\hat{\theta}_B$, for $s < w[1,n]$, while $\hat{\theta}_{MINHS}$ drops at a rate of $2s$. This differing rate of descent explains the generally good performance of $\hat{\theta}_{AVG}$ noted by Dannenbring (1977) and Nydick and Weiss (1994). It also explains Dannenbring's (1977) observations that $\hat{\theta}_{MINHS}$ performs better with a relatively weak heuristic and that the difference in performance between the two decreases as the size of the W sample increases. When s is from a much stronger heuristic than that associated with W, $\hat{\theta}_{MINHS}$ will be much lower than $\hat{\theta}_{AVG}$, resulting in greater bias on the average (due to the increased chance of negative bias). However, as the size of the W sample increases, the difference between the information in this sample and the information in s tends to decrease, and the difference between $\hat{\theta}_{MINHS}$ and $\hat{\theta}_{AVG}$ becomes small.

Appendix B. A Geometric Argument for the Appropriateness of a Weibull Problem Class Model

In linear programming there is a clear graphical means of justifying the Weibull probability distribution as an appropriate model for the behavior of the objective function value, z, as it nears the optimal value θ. Figure 47 depicts the feasible region for a two-variable and a three-variable linear programming problem. The feasible regions are oriented so that the objective function values increase as we travel directly up in the diagram. Five particular solution values are depicted, θ, z_1, z_2, z_3, and z_4. The slice of the feasible region corresponding with any of these particular solution values is called a level set. The level sets are represented in the diagram by dashed gray lines. In the three-dimensional figure, the level sets are actually the triangular region represented by the dashed gray lines. The suitability of the Weibull distribution for modeling the Z random variable is visible in how the size of these level sets changes as z approaches θ. As seen from Figure 47 as long as z is "near" θ, the size of the level set associated with z varies in proportion to $(z - \theta)^{D-1}$, where D = the dimensionality. In this context, z is "near" θ where there are exactly D binding hyperplanes bounding the set of feasible solutions in its level set. Note that the size of the level set of z is a direct reflection of $f(z) \equiv P$(a feasible solution has solution value z).

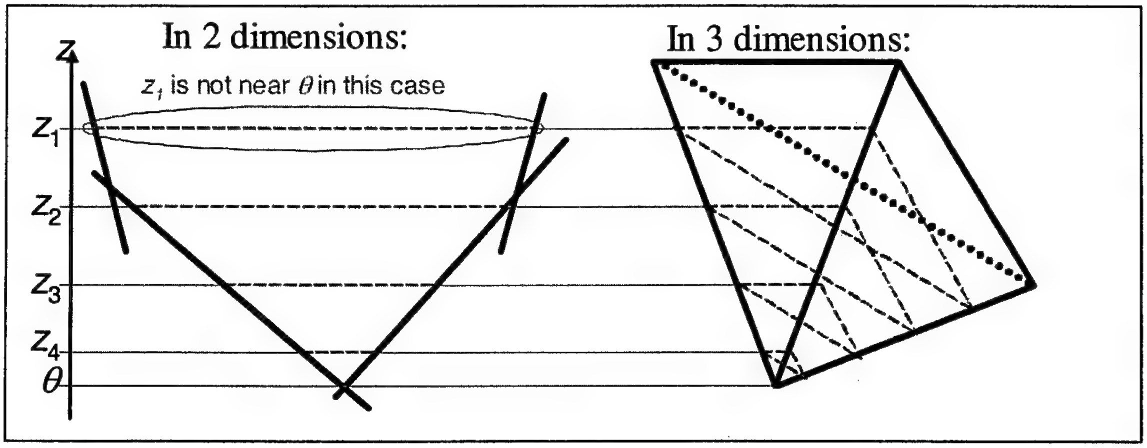

Figure 47. Optimal-Value Diagram for
Continuous-Variable Optimization

Recall that the Weibull(θ,α,β) probability distribution has density function $f(z) = \alpha\beta^{-\alpha}(z-\theta)^{\alpha-1}e^{-((z-\theta)/\beta)^{\alpha}}$, for $z \in (\theta, \infty)$. It's immediately apparent that the $(z-\theta)^{\alpha-1}$ term is appropriate for modeling the probability of a solution value z while z is "near" θ as discussed above. The $\exp(-((z - \theta)/\beta)^{\alpha})$ term handles z distant from θ by causing the density to tail off for very large values of z. The final term, $\beta\alpha^{-\alpha}$, is just the normalization constant to ensure that the density function will integrate to 1 over (θ, ∞).

Appendix C. The Distribution of Minima of Weibull Random Variables

Suppose X has a Weibull distribution with location parameter θ, shape parameter α, and scale parameter β. What is the distribution of $X[1,n]$, the minima of a sample of size n on X?

$$
\begin{aligned}
F_{X[1,n]}(x) &= P(X[1,n] \leq x) \\
&= 1 - P(X[1,n] > x) \\
&= 1 - P\{(X(1) > x) \bigcap (X(2) > x) \bigcap \cdots \bigcap (X(n) > x)\} \\
&= 1 - \prod_{i=1}^{n} P\{X(i) > x\}, \text{ by independence} \\
&= 1 - \prod_{i=1}^{n} (1 - P\{X(i) \leq x\}) \\
&= 1 - (1 - F_X(x))^n, \text{ by identically distributed} \\
&= 1 - \left(1 - \left(1 - \exp\left\{-\left(\frac{x-\theta}{\beta}\right)^{\alpha}\right\}\right)\right)^{n}, \text{ since X\textasciitilde Weibull}(\theta, \alpha, \beta) \\
&= 1 - \left(\exp\left\{-\left(\frac{x-\theta}{\beta}\right)^{\alpha}\right\}\right)^{n} \\
&= 1 - \left(\exp\left\{-n\left(\frac{x-\theta}{\beta}\right)^{\alpha}\right\}\right) \\
&= 1 - \left(\exp\left\{-\left(n^{1/\alpha}\right)^{\alpha}\left(\frac{x-\theta}{\beta}\right)^{\alpha}\right\}\right) \\
&= 1 - \left(\exp\left\{-\left(\frac{1}{n^{-1/\alpha}}\right)^{\alpha}\left(\frac{x-\theta}{\beta}\right)^{\alpha}\right\}\right) \\
&= 1 - \left(\exp\left\{-\left(\frac{x-\theta}{\beta n^{-1/\alpha}}\right)^{\alpha}\right\}\right), \text{ which is Weibull}\left(\theta, \alpha, \beta n^{-1/\alpha}\right).
\end{aligned}
$$

Appendix D. Enumerating Feasible Solution Values for Small Euclidean TSPs

In order to determine the feasibility of using a standard continuous probability model for the problem class distribution, we enumerated the feasible solution values to a collection of 6-node and 9-node Euclidean TSPs. In total we explored 10 6-node instances and 10 9-node instances.

All problem instances we considered were generated by randomly drawing independent (x,y) coordinates for each node with X~Uniform[0,1] and Y~Uniform[0,1]. Euclidean distances were then calculated between each node in a particular problem instance. Feasible solution values were enumerated by systematically generating all permutations of the nodes from a fixed initial node.

Since we expect the minisum and minimax objective function versions of the problem to result in different problem class solution value models, we calculated both the sum of distances and the longest arc distance for each feasible solution. We plotted histograms and empirical CDFs for each of the 50 design points (25 Euclidean TSP instances under each of the two objective functions).

Figure 48 and Figure 50 show that the solution-value histograms for the minisum problem instances are generally bell-shaped. Moreover, Figure even at the level of individual problem instance, 9-node instances with minisum objective are fairly smooth. This supports the idea of a continuous probability model for the problem class distribution. In fact, the repeated bell-shape suggests a central limit theorem is coming into play. As a result, a normal distribution may be a good candidate for problem classes where the objective function is a sum of many real numbers.

In contrast, Figure 49 and Figure 51 show that the minimax objective results in solution-value histograms that are shaped more like an exponential distribution. These histograms are not as smooth as those in Figure 48 and Figure 50, since the minimax objective function leads to more discrete behavior as many distinct solution vectors may contain the same longest arc. The histograms can be expected to become increasingly

smooth as the number of nodes increases. However, this smoothing effect will be markedly slower than in the case of the minisum objective problem class.

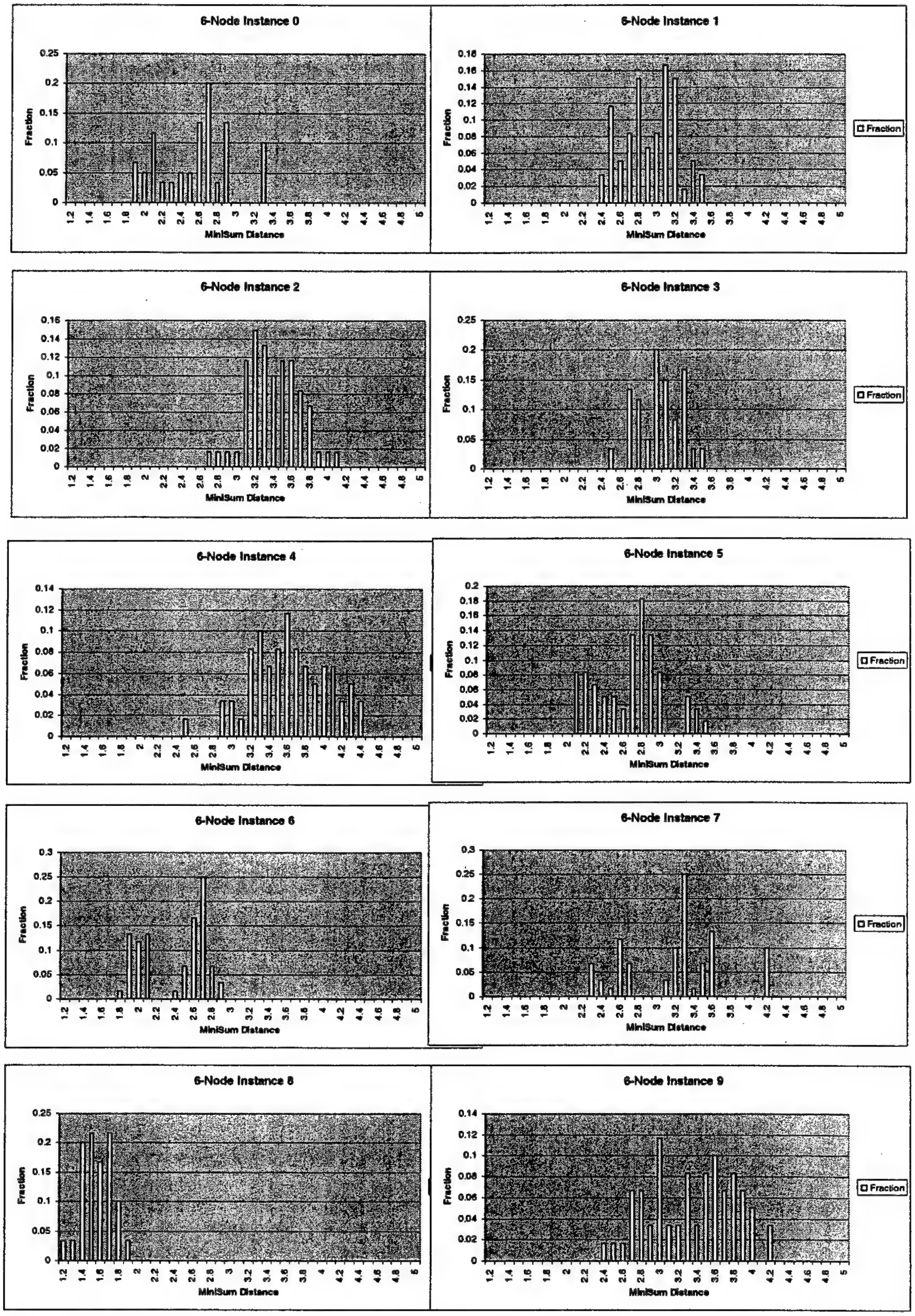

Figure 48. Solution-Value Histograms for 6-Node Minisum Euclidean TSP Instances

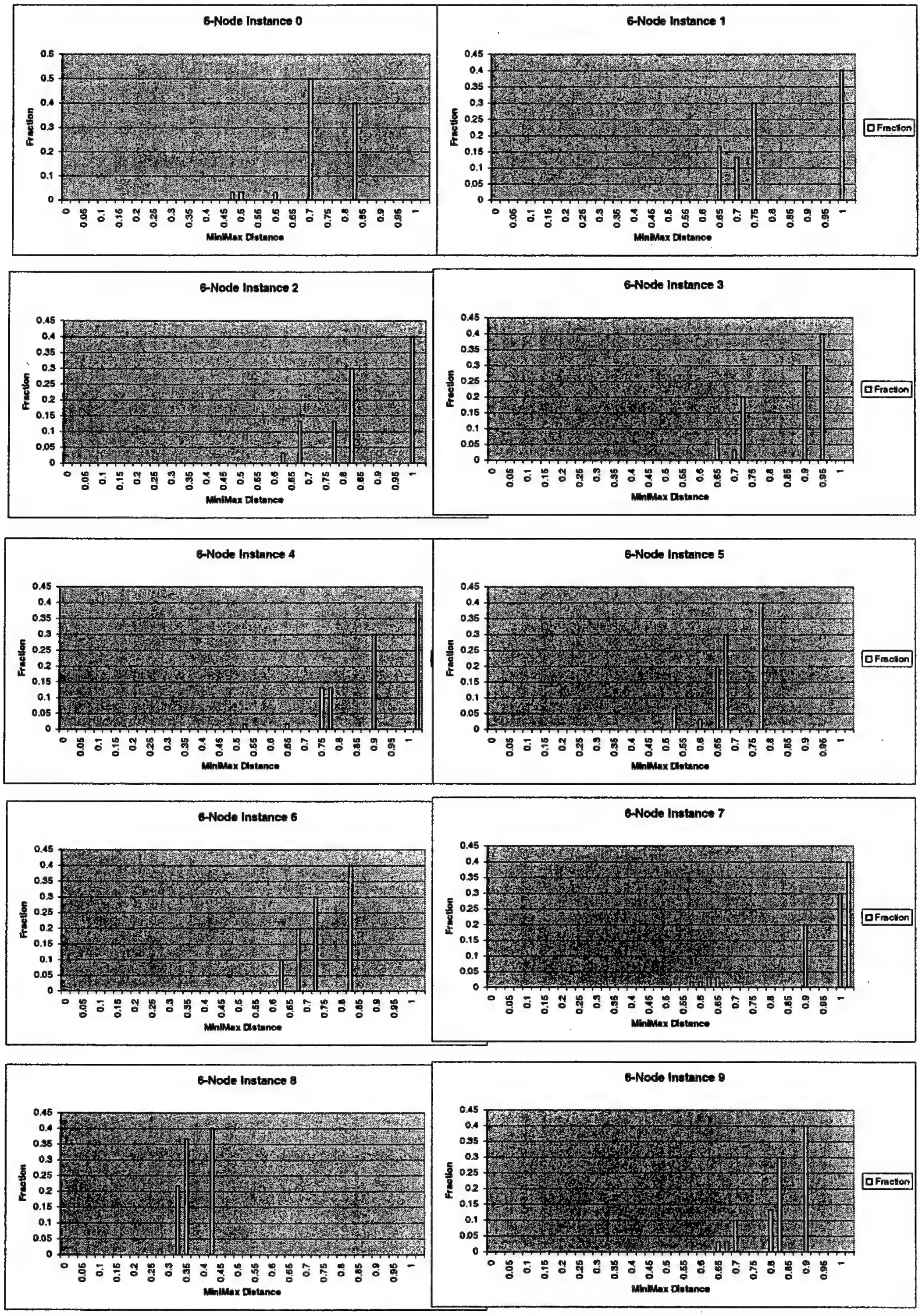

Figure 49. Solution-Value Histograms for 6-Node
Minimax Euclidean TSP Instances

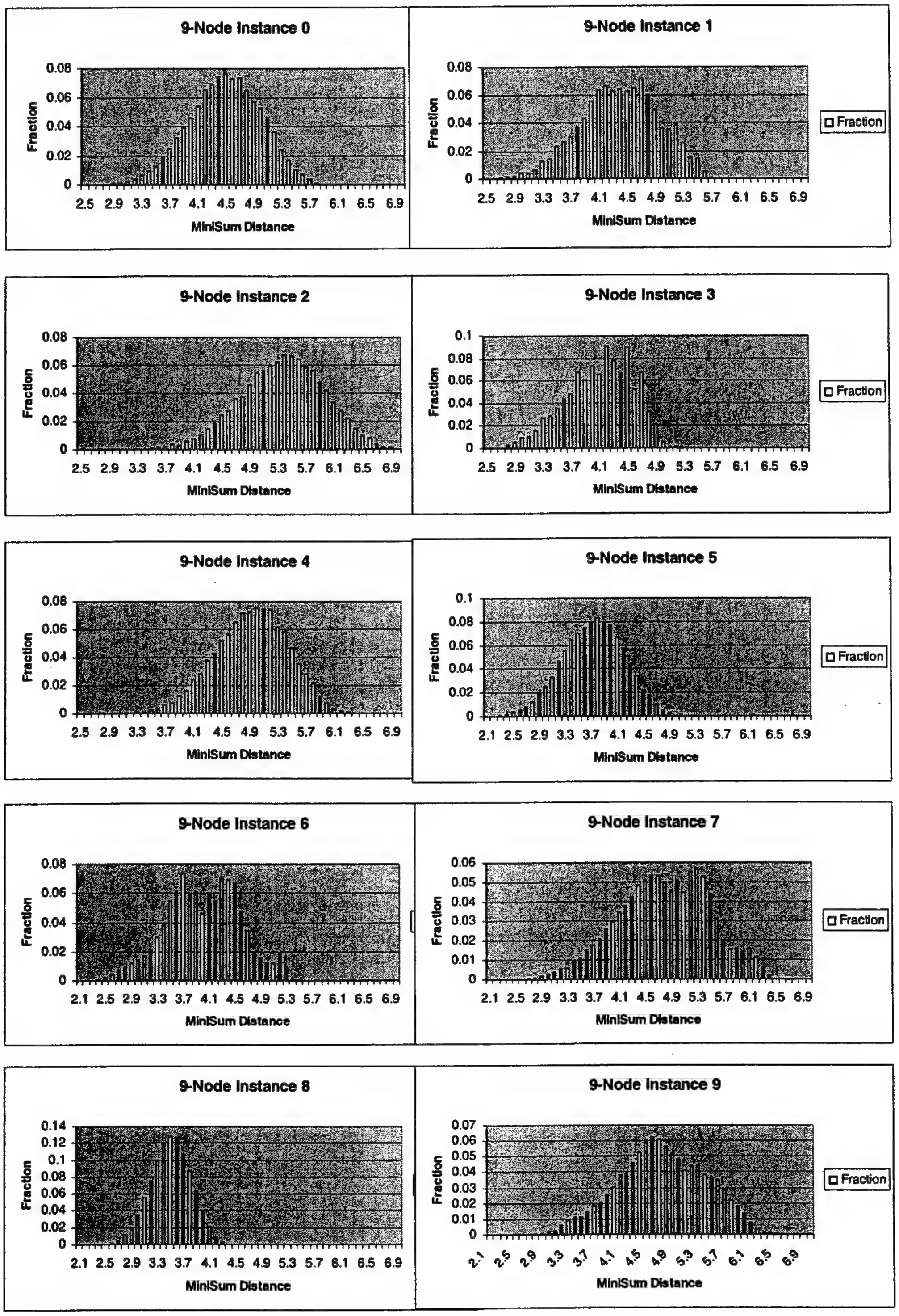

Figure 50. Solution-Value Histograms for 9-Node Minisum Euclidean TSP Instances

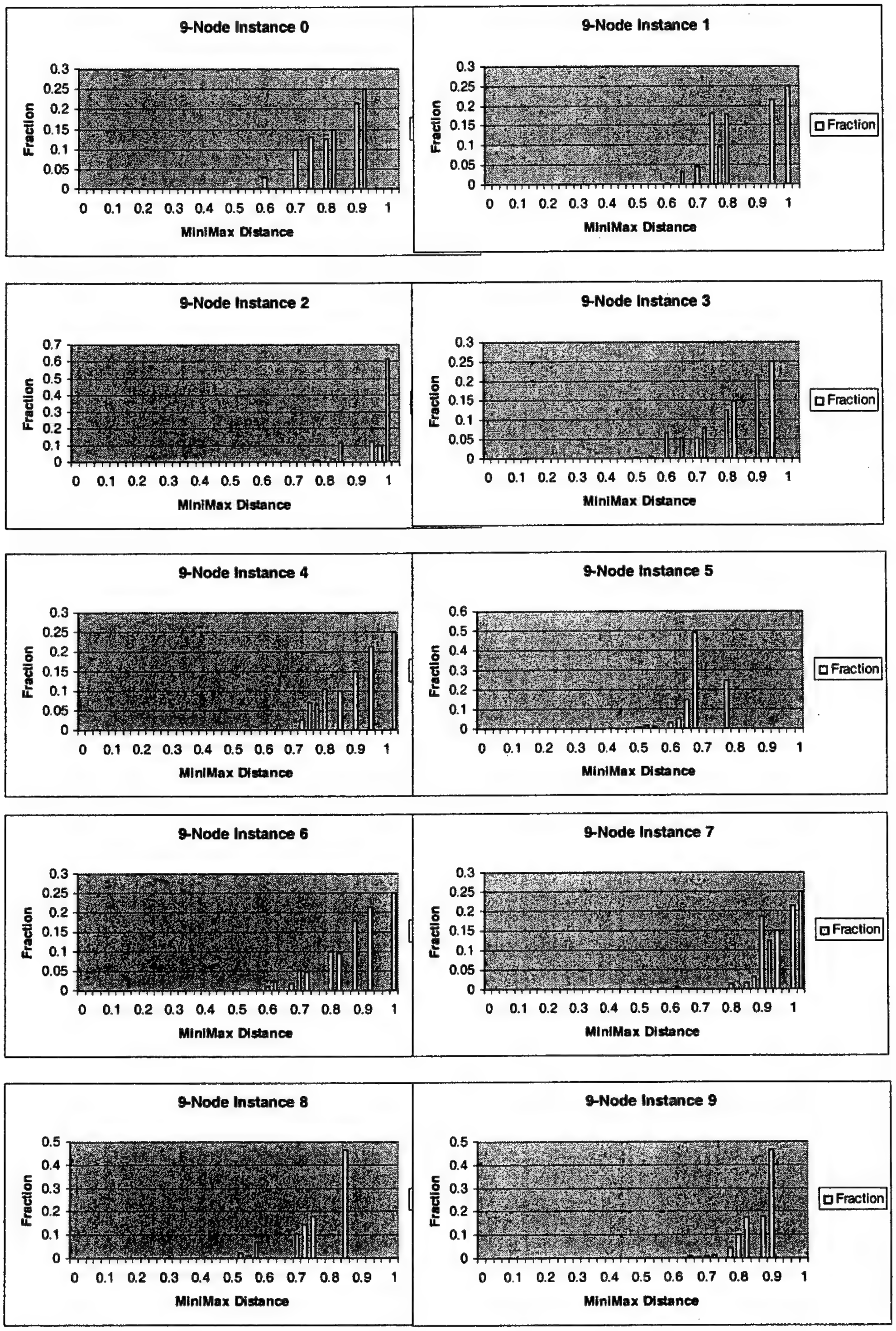

Figure 51. Solution-Value Histograms for 9-Node Minimax Euclidean TSP Instances

Appendix E. C Code for Bayesian Update of P-H Probability Model with k=2

In order to allow the interested reader to replicate or advance upon the empirical work contained in this document, we are enclosing a copy of all relevant computer code. This code was written in C/C+ and run on PCs using the Microsoft Visual Studio compiler.

```
/* Updatek4.c, Angela Giddings, 08 Jun 02*/
/*  Perform Bayesian update on P-H Likelihood Model for k = 4 distinct solution value observations */

#include <stdio.h>
#include <string.h>
#include <stdlib.h>
#include <math.h>

/**********************************************************************
Global Variables:
**********************************************************************/
int psiType, zType, pType, eta;
double psiL, psiU, psiMu, psiSigma,
        zL, zU, zMu, zMuMu, zMuSigma, zSigma, ksee, etaL, etaU;
double *p;
char heurfile[100], outfile[100];
int num, numbins, numcolumns, column, psi, row;
double *theta, *postnum;
double *w, *ct;

/**********************************************************************
Functions:
**********************************************************************/
void main( void );

/* get prior parameters, number of discrete heuristic solution values
to expect, and heuristic file name */
int get_prior( void );

/* get heuristic solution value data */
int get_data( void );

/* output posterior numerator to a file */
void postout( double eta, double zMu, double zSigma, double theta, double postnum, int option );

/* calculate approximate probabilities for Normal CDF */
double norm2(double x, double mean, double sd, int cum);
    /* cum = 0 returns pdf probabilities */
```

```
/*******************************************************************/
/*   Function: main                                                */
/*******************************************************************/
void main(void)
{
    int i, t1, t2, t3, t4, gridpt, option,
        zMuDraw, zSigmaDraw, numzMuDraws, numzSigmaDraws,
        etaDraw, numetaDraws;
    double binsize, spikeProd, curTerm, z1Term, Product, eta, zMu0, zSigma0, zSigmaU;
    double cdfL, cdfU, pdfTheta, cdfTheta, ncoef;
    double *pdf, *cdf;

    if ( get_prior() == 0 )
    {
        /* allocate space for vectors */
        theta = ( double * ) malloc( (numbins + 2)*( sizeof(double) ) );
        postnum = ( double * ) malloc( (numbins + 2)*( sizeof(double) ) );
        w = ( double * ) malloc( num * ( sizeof(double) ) );
        ct = ( double * ) malloc( num * ( sizeof(double) ) );
        pdf = ( double * ) malloc( num * ( sizeof(double) ) );
        cdf = ( double * ) malloc( num * ( sizeof(double) ) );

        if (get_data() == 0)
        {
            /* divide range on theta into numbins subintervals of size binsize */
            binsize = ( w[1] - zL ) / numbins;

            /* printf("\nbinsize = %g", binsize); */

            /* initialize output file */
            option = 0;
            postout( 0, 0.0, 0.0, 0.0, 0.0, option );

            if ( psiType <= 1 )
            {
                /* allocate space for p vector */
                p = ( double * ) malloc( psi * (sizeof(double)) );

                if ( zType <= 4 )
                {
                    numzMuDraws = 0;
                    numzSigmaDraws = 0;
                    zMu0 = zMuMu;
                    zSigma0 = 0.1;
                    numcolumns = numzMuDraws*numzSigmaDraws;
                    row = 0;
```

```
if ( pType == 2 )
     numetaDraws = 12;
else numetaDraws = 0;

printf( "\n Number of Eta draws: %d", numetaDraws );

/* Draw eta */
for ( etaDraw = 0; etaDraw <= numetaDraws; etaDraw ++ )
{
     eta = etaL + (etaU-etaL) * ( (1.0*etaDraw)/(1.0*numetaDraws) );

     printf( "\n eta: %g", eta );

     /* Get p-vector */
     for ( i = 1; i <= psi; i ++ )
     {
          p[i] = pow( ( ((psi - i +1)*1.0)/(psi*1.0) ),eta )
               - pow( ( ((psi-i)*1.0)/(psi*1.0) ),eta );
          /*
          printf("\np%d = %g ", i, p[i]);
          */
     }

/* Draw zMu and zSigma */
for ( zMuDraw = 0; zMuDraw <= numzMuDraws; zMuDraw ++ )
{
for ( zSigmaDraw = 0; zSigmaDraw <= numzSigmaDraws; zSigmaDraw ++ )
{
     zMu = zL + (zU-zL) * ( (1.0*zMuDraw)/(1.0*numzMuDraws) );

     zSigma = zSigma0 +
          (zSigmaU-zSigma0) * ( (1.0*zSigmaDraw)/(1.0*numzSigmaDraws) );

     /* Use following two lines (instead of previous)
     if problem class hyperparameters fixed */
     /*
     zMu = zMuMu;
     zSigma = zMuSigma;
     */

     cdfL = norm2( zL, zMu, zSigma, 1 );
     cdfU = norm2( zU, zMu, zSigma, 1 );
     ncoef = cdfU - cdfL;

     /*
     printf("\nDraw %d %d zMu=%g zSigma=%g",
          zMuDraw, zSigmaDraw, zMu, zSigma);
     printf("\n F(L=%g)=%g F(U=%g)=%g ncoef=%g",
          zL, cdfL, zU, cdfU, ncoef );
     */
```

```
/* calculate (untruncated) normal pdf, and cdf for each w[i] */
for ( i = 1; i < num+1; i++)
{
    cdf[i] = norm2( w[i], zMu, zSigma, 1 );
    pdf[i] = norm2( w[i], zMu, zSigma, 0 );
}

/* Calculate continuous part of posterior */

for ( gridpt = 0; gridpt < numbins; gridpt ++ )
{
    row ++;
    theta[gridpt] = zL+gridpt*binsize;
    postnum[gridpt] = 0.0;
    cdfTheta = norm2( theta[gridpt], zMu, zSigma, 1 );
    pdfTheta = norm2( theta[gridpt], zMu, zSigma, 0 );

    /*
    printf("\nCalculating gridpt %d: theta=%g F(theta)=%g f(theta)=%g",
        gridpt, theta[gridpt], cdfTheta, pdfTheta);
    */
    /* z1Term = f(theta[gridpt]) */
    z1Term = pdfTheta/ncoef;

    /*
    printf("\n i=1 f(%g)~=%g", theta[gridpt], z1Term);
    */

    t2 = 3;
    t3 = 4;
    t4 = 5;
    for ( t1 = 2; t1 < t2; t1 ++ )
    {   /* locate w[1] as some z[i] for this term of sum */
        /*
        printf("\nt1=%d", t1);
        */
    for ( t2 = t1 + 1; t2 < t3; t2 ++ )
    {   /* locate w[2] as some z[i]>w[1] for this term of sum */
        /*
        printf("\nt2=%d", t2);
        */
    for ( t3 = t2 + 1; t3 < t4; t3 ++ )
    {   /* locate w[3] as some z[j] w/ j > i for this term of sum */
        /*
        printf("\nt3=%d", t3);
        */
    for ( t4 = t3 + 1; t4 < psi + 1; t4 ++ )
    {   /* locate w[4] as some z[h] w/ h > j for this term of sum */
        /*
        printf("\nt4=%d", t4);
        */
```

```
Product = 10E10 * z1Term;
/* 10E10 is to help avoid underflow error */
/*
printf("\nProduct=%g initially", Product);
*/
for ( i = 2; i < psi + 1; i ++ )
{
    if ( i < t1 )
    {
        /* curTerm = ( F(w[1]) - F(theta[gridpt]) ) */
        curTerm = (cdf[1] - cdfTheta)/ncoef;
        /*
        printf("\n i=%d F(w[1])-F(%g)~=%g",
                i, theta[gridpt], curTerm);
        */
    }
    if ( i == t1 )
    {
        /* curTerm = f(w[1]) pow(p[t], ct[1]) */
        curTerm = pow(p[t1], ct[1]) * pdf[1]/ncoef;
        /*
        printf("\n i=%d p[%d]^%g f(w[1])~=%g",
                i, t1, ct[1], curTerm);
        */
    }
    if ( (i > t1) && (i < t2) )
    {
        /* curTerm = ( F(w[2]) - F(w[1]) ) */
        curTerm = (cdf[2] - cdf[1])/ncoef;
        /*
        printf("\n i=%d F(w[2])-F(w[1])~=%g", i, curTerm);
        */
    }
    if ( i == t2 )
    {
        /* curTerm = f(w[2]) pow(p[tp], ct[2]) */
        curTerm = pow(p[t2], ct[2]) * pdf[2]/ncoef;
        /*
        printf("\n i=%d p[%d]^%g f(w[2])~=%g",
                i, t2, ct[2], curTerm);
        */
    }
    if ( (i > t2) && (i < t3) )
    {
        /* curTerm = ( F(w[3]) - F(w[2]) ) */
        curTerm = (cdf[3] - cdf[2])/ncoef;
        /*
        printf("\n i=%d F(w[3])-F(w[2])~=%g", i, curTerm);
        */
    }
```

```
                    if ( i == t3 )
                    {
                        /* curTerm = f(w[3]) pow(p[t3], ct[3]) */
                        curTerm = pow(p[t3], ct[3]) * pdf[3]/ncoef;
                        /*
                        printf("\n i=%d p[%d]^%g f(w[3])~=%g",
                                i, t3, ct[3], curTerm);
                        */
                    }
                    if ( (i > t3) && (i < t4) )
                    {
                        /* curTerm = ( F(w[4]) - F(w[3]) ) */
                        curTerm = (cdf[4] - cdf[3])/ncoef;
                        /*
                        printf("\n i=%d F(w[4])-F(w[3])~=%g", i, curTerm);
                        */
                    }
                    if ( i == t4 )
                    {
                        /* curTerm = f(w[4]) pow(p[t4], ct[4]) */
                        curTerm = pow(p[t4], ct[4]) * pdf[4]/ncoef;
                        /*
                        printf("\n i=%d p[%d]^%g f(w[4])~=%g",
                                i, t4, ct[4], curTerm);
                        */
                    }
                    if ( i > t4 )
                    {
                        /* curTerm = ( 1 - F(w[4]) ) */
                        curTerm = (cdfU - cdf[4])/ncoef;
                        /*
                        printf("\n i=%d 1-F(w[4])~=%g", i, curTerm);
                        */
                    }

                    /* (Product * curTerm) */
                    Product = Product * curTerm;
                    /*
                    printf("\n Product=%g", Product);
                    */
                } /* for i < psi + 1 in calculating continuous part */

                postnum[gridpt] = postnum[gridpt] + Product;
                /*
                printf("\n theta[%d]=%g postnum=%g",
                        gridpt, theta[gridpt], postnum[gridpt]);
                */

            } /* for t4 in calculating continuous part */
            } /* for t3 in calculating continuous part */
            } /* for t2 in calculating continuous part */
            } /* for t1 in calculating continuous part */
```

```
        option = 2;
        postout( eta, zMu, zSigma, theta[gridpt], postnum[gridpt], option );
    } /* for gridpt in calculating continuous part */

    /* Calculate discrete part of posterior */
    row ++;
    theta[numbins] = w[1];
    postnum[numbins] = 0.0;

    t2 = 2;
    t3 = 3;
    t4 = 4;
    for ( t2 = 2; t2 < t3; t2 ++ )
    {   /* locate w[2] among the z[i]>w[1]=z[1] */

    for ( t3 = t2 + 1; t3 < psi + 1; t3 ++ )
    {   /* locate w[3] among the z[j]>z[i]>w[1]=z[1] */

    for ( t4 = t3 + 1; t4 < psi + 1; t4 ++ )
    {   /* locate w[4] among the z[h]>z[j]>z[i]>w[1]=z[1] */
        /*
        printf("\nCalculating spike: t2=%d t3=%d t4=%d", t2, t3, t4);
        */
        /* curTerm = f(w[1]) pow(p[1],ct[1]) */
        z1Term = pow(p[1], ct[1]) * pdf[1]/ncoef;
        spikeProd = 10E10 * z1Term;
        /* 10E10 is to help avoid underflow error */

        /*
        printf("\ni=1 p[1]^%g f(w[1]=z[1])~=%g", ct[1], z1Term);
        */

        for ( i = 2; i < psi + 1; i ++ )
        {
            if ( i == t2 )
            {
                /* curTerm = f(w[2]) pow(p[t],ct[2]) */
                curTerm = pow(p[t2], ct[2]) * pdf[2]/ncoef;
                /*
                printf("\ni=%d p[%d]^%g f(w[2])~=%g",
                    i, t2, ct[2], curTerm);
                */
            }
            if ( i < t2 )
            {
                /* curTerm = ( F(w[2])-F(w[1]) ) */
                curTerm = (cdf[2] - cdf[1])/ncoef;
                /*
                printf("\ni=%d F(w[2])-F(w[1])~=%g", i, curTerm);
                */
            }
```

```
        if ( (i > t2) && (i < t3) )
        {
            /* curTerm = ( F(w[3]) - F(w[2]) ) */
            curTerm = (cdf[3] - cdf[2])/ncoef;
            /*
            printf("\n i=%d F(w[3])-F(w[2])~=%g", i, curTerm);
            */
        }
        if ( i == t3 )
        {
            /* curTerm = f(w[3]) pow(p[t3], ct[3]) */
            curTerm = pow(p[t3], ct[3]) * pdf[3]/ncoef;
            /*
            printf("\n i=%d p[%d]^%g f(w[3])~=%g",
                    i, t3, ct[3], curTerm);
            */
        }
        if ( (i > t3) && (i < t4) )
        {
            /* curTerm = ( F(w[4]) - F(w[3]) ) */
            curTerm = (cdf[4] - cdf[3])/ncoef;
            /*
            printf("\n i=%d F(w[4])-F(w[3])~=%g", i, curTerm);
            */
        }
        if ( i == t4 )
        {
            /* curTerm = f(w[4]) pow(p[t4], ct[4]) */
            curTerm = pow(p[t4], ct[4]) * pdf[4]/ncoef;
            /*
            printf("\n i=%d p[%d]^%g f(w[4])~=%g",
                    i, t4, ct[4], curTerm);
            */
        }
        if ( i > t4 )
        {
            /* curTerm = ( 1 - F(w[4]) ) */
            curTerm = (cdfU - cdf[4])/ncoef;
            /*
            printf("\ni=%d 1-F(w[4])~=%g", i, curTerm);
            */
        }

        spikeProd = spikeProd*curTerm;

    } /* for i < psi + 1 in calculating discrete part */

    postnum[numbins] = postnum[numbins] + spikeProd;
    /*
    printf("\nSpike: theta[%d]=w[1]=%g postnum=%g",
        numbins, theta[numbins], postnum[numbins]);
    */
```

```
                    } /* for t4 < psi + 1 in calculating discrete part */
                    } /* for t3 < t4 in calculating discrete part */
                    } /* for t2 < t3 in calculating discrete part */

                    option = 2;
                    postout( eta, zMu, zSigma, w[1], postnum[numbins], option );

                option = 1;
                postout( eta, zMu, zSigma, 0.0, 0.0, option );

                }}} /* for etaDraw, zMuDraw, and zSigmaDraw */

            } /* if zType=4 */
        } /* if psiType=1 */

    } /* if get data */
free( theta );
} /* if get prior */

printf("\n FINISHED");

return;
}
```

```
/*********************************************************************/
/*   Function: get_prior                                             */
/*   Reads in a file containing parameters of the prior distribution */
/*   and the number of heuristic runs and estimation subintervals    */
/*  that will be used.                                               */
/*********************************************************************/

int get_prior( void )
/* returns 0 if ok, 1 OW */
{
    char in_file[100];
    FILE *in;
    char name[100];
    int psiType, zType;

    /* open the parameter file */
    printf( "\nEnter the parameter file name: " );
    scanf( " %s", in_file );
    in = fopen( in_file, "r" );
    if( in == NULL )
    {
        printf( "Parameter error: parameter file does not exist.\n" );
        return 1;
    }

  /* line 1: instance name */
  fgets( name, 100, in );
  printf( "Name: %s", name );

  /* line 2: type of prior on psi, type of prior on z, type of heuristic & eta value
    for psi or z:
        1=fixed value, 2=uniform,
        3=doubly truncated normal, 4=doubly truncated normal w/ hyperparameter prior
    for heuristic:
        0 ksee etaL etaL= known that heuristic returns best of sample of size etaL
        2 ksee etaL etaU = unknown but thought to be like returning best of eta on [etaL, etaU],
            prior sample size ksee
  */

    fscanf( in, "%d %d %d %lf %lf %lf", &psiType, &zType, &pType, &ksee, &etaL, &etaU );
    printf("\nPrior type %d for psi %d for z %d for p ksee=%g etaL=%g etaU=%g",
        psiType, zType, pType, ksee, etaL, etaU );
```

```
/* line 3: prior parameter(s) for psi */
    if ( psiType == 1 )
    {
        fscanf( in, "%d", &psi );
        printf( "\nPsi = %d", psi );
        if ( psi <= 0 )
        {
            printf( "Psi Parameter error: Must be positive number.\n" );
            return 1;
        }
    }
    else if ( psiType == 2 )
    {
        fscanf( in, "%lf %lf", &psiL, &psiU );
        if ( psiL >= psiU )
        {
            printf( "Psi Parameter error: Upper bound must be greater than lower bound.\n" );
            return 1;
        }
        if ( psiL <= 0 )
        {
            printf( "Psi Parameter error: Lower bound must be positive.\n" );
            return 1;
        }
    }
    else if ( psiType == 3 )
    {
        fscanf( in, "%lf %lf %lf %lf", &psiL, &psiU, &psiMu, &psiSigma );
        printf( "\nPsi is Normal with L=%g U=%g Mu=%g Sigma=%g",
                psiL, psiU, psiMu, psiSigma );
        if ( psiL >= psiU )
        {
            printf( "Psi Parameter error: Upper bound must be greater than lower bound.\n" );
            return 1;
        }
        if ( psiL <= 0 )
        {
            printf( "Psi Parameter error: Lower bound must be positive.\n" );
            return 1;
        }
        if ( (psiMu <= psiL) || (psiMu >= psiU) )
        {
            printf( "Psi Parameter error: Mu must be  between lower and upper bounds.\n" );
            return 1;
        }
    }
    else
    {
        printf( "\n Incorrect prior specification for Psi. \n");
        return 1;
    }
```

```
    /* line 5: number of distinct heuristic runs,
                    number of histogram bins desired */
    fscanf( in, "%d %d %d", &num, &numcolumns, &numbins );
    if ( num <= 0 )
    {
        printf( "Parameter error: Number of runs must be positive.\n" );
        return 1;
    }
    else
    {
        /* line 6: name of file containing (weak) heuristic runs */
        fscanf( in, "%s", heurfile );
    }

    fscanf( in, "%s", outfile );

    printf("\nNum = %d HeurFile = %s OutFile = %s", num, heurfile, outfile);

    fclose( in );

    return 0;
}
```

```
/*******************************************************************/
/*   Function: get_data                                            */
/*   Reads in a file containing num heuristic solution values.     */
/*******************************************************************/

int get_data( void )
/* returns 0 if ok, 1 OW */
{
    int i;
    FILE *in;

    /* open the parameter file */
    in = fopen( heurfile, "r" );
    if( in == NULL )
    {
        printf( "Parameter error: heuristic file does not exist.\n" );
        return 1;
    }

    /* line 1+: distinct heuristic solution value, number of occurences in sample */
    /* Note:  The solution values should be provided in increasing order of sol value
    magnitude */

    for ( i = 1; i <= num; i ++ )
    {
        fscanf( in, "%lf %lf", &w[i], &ct[i] );
        printf("\nIndex %d Soln Value: %g Count %g",
                    i, w[i], ct[i] );
    }

    fclose( in );

    return 0;
}
```

```
/*****************************************************************/
/* Function: norm2
/*    computes the cumulative distribution function P(Y <= y)
*     of a standard normal random variable Y=(X-mean)/sd.
*   input
*     x: normal value
*   output
*     cdf: P(X <= x)
*     pdf: standard normal density at x
*   reference:  C. Hastings, Approximations for Digital Computers.
*     Princeton University Press, Princeton, NJ, 1955.
*   This code is a minor modification, by Bruce Schmeiser, of the code in the IBM
*   scientific subroutine package, 1967, page 78, modified again by Angela Giddings
*   for use as a function and for any normal(mean, sd) by standardizing
*/
double norm2(double x, double mean, double sd, int cum)
{
  double t,y,ay, cdf, pdf;

  y = (x-mean)/sd;
  ay = (float) fabs((double) y);
  t = 1.0 / (ay * .2316419 + 1.0);
  pdf = (float) exp( -(double)(y * y * .5) ) * .3989423;

  if ( cum > 0 )
  {
       cdf = 1.0 - pdf * t * ((((t * 1.330274 - 1.821256) * t
              + 1.781478) * t - .3565638) * t + .3193815);
       if (y < 0.0) cdf = 1 - cdf;
       return cdf;
  }
  else return pdf/sd;

}
```

```
/*******************************************************************/
/* Function: postout                                               */
/* Output posterior to a file.                                     */
/*******************************************************************/
void postout( double eta, double zm, double zsig, double theta, double pnum, int option )
{
  FILE *out;
     if( option == 0 )
     {
          /* open output file initially */
          out = fopen( outfile, "w" );
          if( out == NULL )
          {
               printf( "Parameter error: output file cannot be opened\n" );
               return;
          }
          /* print header row */
          fprintf( out, "eta   zMu      zSigma  Theta     PosteriorNumerator" );
          fprintf( out, "\n" );
          fclose( out );
          return;
     }
     else if( option == 1 )
     {
          /* open output file to append */
          out = fopen( outfile, "a" );
       if( out == NULL )
       {
         printf( "Parameter error: output file cannot be opened\n" );
         return;
       }
       else
       {
          /* print header */
               fprintf( out, "\n" );
               fclose( out );
               return;
       }
     }
```

```
    else (if option > 1)
    {
        /* open output file to append */
        out = fopen( outfile, "a" );
      if( out == NULL )
      {
        printf( "Parameter error: output file cannot be opened\n" );
        return;
      }
      else
      {
        /* print data row */
            fprintf( out, "\n%1.9g  %3.9g  %3.9g  %3.9g  %3.9g ",
                eta, zm, zsig, theta, pnum );
            fclose( out );
            return;
      }
    }
}
```